Qualico-94

2nd International Conference on Quantitative Linquistics

September 20-24 1994

Moscow, Lomonosov Moscow State University
Philological Faculty

2-я Международная конференция по квантитативной лингвистике

20-24 сентября 1994 года

Москва, Московский Государственный университет им. М.В. Ломоносова
Филологический факультет

2-я Международная конференция по квантитативной лингвистике

Lomonosov Moscow State University
Philological Faculty

Trier University

International Informatization Academy

2nd International Conference on Quantitative Linquistics

September 20-24 1994

Moscow, Lomonosov Moscow State University
Philological Faculty

Abstracts of papers

Edited by Anatoliy A. Polikarpov

Moscow - 1994

Московский государственный университет имени М.В. Ломоносова
Филологический факультет

Трирский университет

Международная Академия Информатизации

2-я Международная конференция по квантитативной лингвистике

20-24 сентября 1994 года

Москва, Московский Государственный университет
им. М.В. Ломоносова
Филологический факультет

Тезисы докладов

под редакцией А.А. Поликарпова

Москва - 1994

От оргкомитета

Qualico-94 является второй по счету международной конференцией по квантитативной лингвистике. Первая была проведена в сентябре 1991 года в Трирском университете (Германия).

Настоящая конференция организована по решению Постоянного комитета по проведению конференций по квантитативной лингвистике (Р. Келер, Ш. Эмблтон, А.А. Поликарпов). Она проводится Московским университетом совместно с Трирским университетом в рамках договора о научном и учебном сотрудничестве.

Важную помощь в проведении конференции оказала МеждународнаяАкадемия Информатизации.

Мы признательны за помощь всем тем, кто принял участие в конференции, кто тем или иным образом способствовал ее проведению.

From Organizing Committee

Qualico-94 is the second international conference on Quantitative Linguistics. The first one was held in September 1991 by Trier University.

Qualico-94 was organized according to the decision of Permanent Commitee on holding quantitative linguistics conferences (R. Koelner, S. Embleton, A.A. Polikarpov). It is being held by Lomonosov Moscow State University together with Trier University (Department of Computer Linguistics, professor Reinhard Koehler) within the framework of the Agreement between two Universities on scientific and teaching cooperation.

Important support to Moscow University in converting the idea of the Conference into fact was given by international Informatization Academy.

We acknowledge the help of all those people who participate in it, who was engaged in preparations to it.

The Development of Analytism in the Hungarian Declension.

Tatiana B. Agranat
Lomonosow Moscow State University
Philological Faculty
Russia, 117292, Moscow, Iv.Babushkina 20-13.

Topical paper.

AREA: Structural tendencies in the development of Hungarian language.

Summary:

This paper contains the account of the investigation on the structural tendencies of expression of the nominal declensi on functional-semantic category. One can trace the increase in usage of postposition constructions at the expense of the decrease in usage of case affixes.

I. The tendencies of development of grammatical systems have been studied on the data of the Indoeuropean, Chinese and other languages for a rather long time already. However, Uralic and Altaic languages never yet entered into consideration. This work is an attempt to investigate the history of Hungarian language from this point of view. It is clear that in order to describe the way of development of a language grammar as a whole towards analytism or synthetism it is necessary to consider structural tendencies in expression of functional semantic categories. This talk is an account of the part of a large research on the grammar of Hungarian postpositions, so we shall confine ourselves only to the consideration of nominal declension (in terms of the deep cases), that is to tracing the correlation of usage of case-affixes and postpositions in the history of Hungarian.

II. Precision of the results in the investigation of this kind increases if identical texts written in different epoch are used for data. We have this opportunity, because different parts of the Bible were translated more than once during the Hungarian history, though it is rather difficult to find concurrent parts.

This work is done on the data of Mark and Matthew's gospel translations, made in different historical periods.

1. Munich Codex, year 1466.

2. The first Hungarian printed book, translation of the Gospel by Sylvester Janos.

3. XXth century translation.

III. Constructions with postpositions were taken from each script and the problem whether each of the constructions corresponds to the same (analytical) expression, or these relations are expressed synthetically with the use of a case affix was considered. The investigation of each usage of every postposition produced different results. It is connected with the fact that different postpositions behave in a different way at the length of Hungarian history. The frequency of some of them steadily grows, of others - falls. The use of some increases towards the 16th century, than decreases and vice versa. However, the general tendency is the increase in usage of postpositional construction at the expense of the usage ofcase affixes, as it is seen on the table 1.

The first column in the table is responsible for sunchronic level presented in the script. The second - for number of postpositions in each of the scripts and corresponding to case-affixes at least in one other script. The third - for the number of case-affixes in each of the scripts and correspon-ding to postpositions at least in one other script.

Table 1:

	I	II	III
	XV century	3	17
	XVI century	10	14
	XX century	17	7

IV. The data of the same texts are used for the evaluatuon of the degree of synthetism according to J. Greenberg for each synchronic level. (J. H. Greenberg, A quantitative approach to the morphological typology of language, 'International Journal of American Linguistics', vol. XXVI, #3).

The results of general evaluation are as follows:

XVth century 1.63 M/W
XVIth century 1.47 M/W
XXth century 1.61 M/W

where M stands for morphs and W for words.

The results for evaluation of noun phrases (with the exception of participles) are follows:

XVth century 1,72 M/W
XVIth century 1,61 M/W
XXth century 1,43 M/W

According to the data we may state that maximal degree of analytism corresponds to the 16th century synchronic level, and after it is overcome by the synthetic tendencies. At the background of general drift towards synthetism, we see unquestionable analytical tendencies in the nominal declension of the language. Probably this drift is due to the change in the system of verbal inflexion, where the analytical forms of past tenses have disappeared. It proves once more that general trend in the development of a grammatical system consists of different, sometimes conflicting tendencies modifying ways of expression of various functional semantic categories.

Развитие аналитизма в венгерском склонении.

Агранат Т.В

Резюме:

Настоящий доклад содержит результаты исследования тенденций развития способов выражения именного склонения в венгерском языке. Обнаруживается рост в употреблении постпозитивных конструкций за счет снижения употребления падежных аффиксов.

A Morphological Processor for the Russian Language

Anoshkina Janna G.
Russian language Institute
of the Russian Academy of Sciences
Computer Fund of the Russian language.
Russia, 121019, Moscow, Volkhonka 18/2,
FAX: (095) 201-22-76
E-Mail: irlras@irl.msk.su

Project note.

AREA: Morphology processor for Russian language.

Summary:

The theme of the project note is a morphological processor for the Russian language, in which there are realized both morphological analysis and lemmatization (from the word-form to the dictionary-form) and generation of the word-forms (from the dictionary-form to the word-form).

Two functions: the first - morphological analysis and lemmatization, and the second - generation of the word-forms are realised in the Morhological processor for the Russian language.

A morphological analysis is a procedure in which every word-form receives its morphological information; lemmatization is a procedure in which every word-form receives its dictionary-form (lemma). The dictionary-form for nouns is nominative case singular; for adjectives - the Nominative case singular masculine; for verbs - the Infinitive (including participles and verbal adverbs); for personal pronouns - the Nominative case, for other pronouns - the Nominative case singular; for cardinal numerals - the Nominative case, for ordinal numerals - the Nominative case singular. The morphological information for nouns, adjectives, pronouns, numerals, participles includs gender, number, case; for verbs - aspect, transitivity, tense, person, num-ber.

The result of the generation process is all word-forms of any dictionary-form (each word-form has its morphological information).

The morphological analysis and the generation of the word-forms use the same dictionary base. In the first case we enter into the base through word-forms, in the second - through dictionary-forms. The dictionary base is created from the Grammar Dictionary of the Russian Language by A.A. Zalizniak, it is enriched with proper names and the words that were obtained after the processing of the large amount of texts. Now the dictinary base consists of 80 thousend lemmas, the amount of the word-forms is about 2 million; it occupies 3 Mb. The access speed depends on the caracteristics of the computer, for AT/486 it is 0.12second per word.

All variants of lemmas and all variants of morphological information for a word-form are built after the morphological analisys. In the same way all variants of paradigms for a lemma are built in the process of the generation.

A word-form which was not recognized against the dictionary receives morphological interpretation "by pattern", that is a string or several strings consisting of lemma and morphological information borrowed from the most "similar" word-form. The same is true for the generation: an unrecognized lemma receives several variants of paradigms borrowed from the most "similar" dictionary-form.

The results of the morphological analysis and lemmatization are used for syntactic analysis of texts and for the creation of frequency dictionary-form vocabularies and concordances of texts. The results of the generation are used to enrich the dictionary base with new words and to make easier the manipulation with the automatic concordance.

There is a serving system to support the dictionary base in the actual condition.

Морфологический процессор русского языка

Аношкина Ж.Г.

Резюме:

Тема проектной заметки - описание морфологического процессора для русского языка, в котором реализованы как морфологический анализ и лемматизация (от словоформы к словарной форме), так и образование словоформ (от словарной формы к словоформе)

Vergleich von verschiedenen Methoden der Clusteranalyse in linguistischen Forschungen (Clusteranalyse in der Linguistik)

Mag. phil. Anne Arold
Universität Tartu
Ülikoolistr. 18a, EE2400, Tartu, Estland
Tel. : (372 7) 435 282 Fax: (372 7) 435 440

In den linguistischen Forschungen gewinnt die Clusteranalyse als ein Instrumentarium zum Erkennen von Strukturen in einer Menge von Objekten immer mehr an Bedeutung. Um die Anwendbarkeit verschiedener Verfahren miteinander zu vergleichen, werden in der vorliegenden Untersuchung drei Varianten der Clusteranalyse zur Ermittlung von Klassenzugehörigkeiten innerhalb einer Gruppe von deutschen Adjektiven angewendet:

1. B_k-Methode (*k*-Clusterung);
2. Zentroid-Methode (das statistische Programmsystem BMDP2M);
3. Methode der *k*-Mittelpunkte (das statistische Programmsystem BMDPKM).

1. Allgemeine Prinzipien der Clusteranalyse

Bei allen Varianten der Clusteranalyse handelt es sich um drei Arbeitsetappen:
- Kodierung der Ausgangsdaten, Erstellen der Datenmatrix;
- Ermittlung der Distanz (*d*) (Nähe bzw. Abstand zwischen den Objekten);
- Konstruieren des Cluster-Systems, das die Objekte auf verschiedenen Ebenen (*h*) gruppiert. (Die numerischen Werte von *h* werden jeweils von der Methode und vom entsprechenden Computer-Programm bestimmt.)

Die zwei letztgenannten Etappen werden in der Regel mit Hilfe des Elektronenrechners durchgeführt (Ääremaa, 1981; Tuldava, 1987). Bei der vorliegenden Arbeit wurde dazu der Elektronenrechner EC-1060 des Rechenzentrums der Universität Tartu benutzt.

Aufgrund der Datenmatrix wird der gegenseitige Abstand der Objekte berechnet. Man kann dazu verschiedene Verfahren anwenden:

a) die Distanz d_S aufgrund des Sörensen-Ähnlichkeitskoeffizienten (r_S) (Ääremaa, 1991);

$$d_S(O_a,O_b) = 1 - r_S(O_a,O_b),$$

wobei

$$r_S(O_a,O_b) = \frac{2C}{A+B}$$

A - Anzahl der verzeichneten Merkmale beim Objekt a;
B - Anzahl der verzeichneten Merkmale beim Objekt b;
C - Anzahl der übereinstimmenden Merkmale bei den Objekten a und b;

b) die Distanz d_L aufgrund des Koeffizienten der linearen Korrelation (r_L):

$$d_L(O_a,O_b) = 1 - r_L(O_a,O_b),$$

wobei

$$r_L(O_a,O_b) = \frac{1}{\sigma_a\sigma_b} \sum_{i=1}^{m} (x_{i_a} - \bar{x}_{i_a})(x_{i_b} - \bar{x}_{i_b})$$

σ - Standardabweichung;
m - Anzahl der Merkmale;
x_i - Merkmalswerte;
-
$\bar{x}_i$ - Mittelwerte (Tuldava, 1987);

c) zwei Varianten der sog. *p*-Distanz (Minkowski-Distanz), die in verallgemeinerter Form nach der Formel

$$d_p(O_a, O_b) = \left[\sum_{i=1}^{m} (x_{i_a} - x_{i_b})^p\right]^{\frac{1}{p}}$$

ausgerechnet wird, wobei x_a und x_b Merkmalswerte bezeichnen.
Wir haben zwei Varianten der *p*-Distanz angewendet:
- den euklidischen Abstand d_E ($p = 2$):

$$d_E(O_a, O_b) = \sqrt{(x_a-x_b)^2 + (y_a - y_b)^2}$$

wobei x und y Merkmalswerte bezeichnen (Aivazjan u.a.,1989);
- den Hammingabstand d_H (Hartung / Elpelt, 1986) ($p = 1$), der vereinfacht in der Formel

$$d_H(O_a,O_b) = \sum_{i=1}^{m} \left|(x_{i_a} - x_{i_b})\right|$$

ausgedrückt wird.
Die letztgenannte Variante ist besonders gut anwendbar bei einem Binärcode (Merkmalswerte "1" oder "0") (Aivazjan u.a.,1989).
Aufgrund der berechneten Distanzen erfolgt die Clusterung, d.h. das Konstruieren des Cluster-Systems, wobei zwischen hierarchisch und nicht-hierarchisch organisierter Clusterung unterschieden wird. Das hierarchische Cluster-System kann seinerseits entweder agglomerativ oder divisibel aufgebaut werden (Tuldava, 1987).

1.1. B_k-Methode

Die B_k-Methode ist eine weiterentwickelte Variante der hierarchisch-agglomerativen Methode. Es handelt sich um die *k*-Clusterung, wobei *k* die Zahl der Elemente bezeichnet, deren Überlappen bei den benachbarten Clusters zugelassen wird (Ääremaa, 1978; Tuldava, 1987). Bei $k = 1$ ist das Ergebnis der Clusterung sowie der ganze Verlauf dieses Vorgangs in Form eines Dendrogramms darstellbar. Dieses Cluster-System sieht allerdings einigermaßen entstellt aus, da die fertigen Clusters auf den höheren Ebenen als neue Objekte behandelt werden und mit weiteren Objekten zusammengeschlossen werden können, deren Abstand von einigen Elementen innerhalb dieser Clusters den auf dieser Ebene zur Clusterbildung zugelassenen Höchstwert der Distanz übersteigt. Es besteht die Möglichkeit, diese Entstellungen zu messen und die Ebene zu finden, auf der sie am kleinsten sind (Ääremaa, 1981).

Etwas mehr Interpretationsmöglichkeiten bietet die *k*-Clusterung, wenn $k > 1$. Dabei handelt es sich um die Clusterung mit teilweise überlappenden Clusters (*k* bezeichnet die Zahl der Objekte, die bei zwei benachbarten Clusters übereinstimmen können). Diese Methode ermöglicht es, die verbindenden Elemente zwischen einzelnen Clusters festzustellen (Ääremaa, 1981). Im Rahmen der vorliegenden Untersuchung wurde die 2-Clusterung ausgeführt.

1.2. Zentroid-Methode

Bei der Zentroid-Methode, die sich auf die Hammingdistanz gründet, werden die Clusters nach denselben Prinzipien aufgebaut, aber dabei ist es möglich, die

einzelnen Schritte der Clusterung besser zu unterscheiden, als es bei der einfachen *k*-Clusterung der Fall ist. Nach der Verbindung von zwei Punkten (Objekten) wird hier ein sog. Pseudopunkt berechnet, der das Mittel von Merkmalswerten der entsprechenden Punkte darstellt. Danach werden die Distanzen der potentiellen Cluster-Elemente zum Pseudopunkt berechnet, und dem vorhandenen Cluster wird derjenige Punkt (Objekt) angegliedert, der dem Pseudopunkt am nächsten liegt (Ääremaa, 1991).

1.3. Die Methode der k-Mittelpunkte

Wie die Zentroid-Methode, so ist auch die Methode der *k*-Mittelpunkte eine weiterentwickelte Variante des statistischen Programmsystems BMDP. Bei dieser Methode werden die Objekte in *k* Clusters eingeteilt. Anschließend werden die Mittelpunkte der Clusters sowie die Distanzen aller Objekte von diesen Mittelpunkten berechnet. Auf diese Weise kann man die Struktur jedes Clusters genau feststellen, d.h. zwischen der Kernzone und der Peripherie jedes Clusters unterscheiden. Darüber hinaus kann mit Hilfe dieser Methode auch die Stellung der einzelnen Clusters in der gesamten Objektmenge ermittelt werden. Der größte Vorteil der genannten Methode besteht aber darin, daß die Clusters hier auch qualitativ beschrieben werden. Von jedem Cluster wird ein Profil zusammengestellt, auf dem die Mittelwerte aller Merkmale (samt ihren Standardabweichungen in dem jeweiligen Cluster) aufgezeichnet sind. Aufgrund dieser Profildarstellungen kann man erfahren, durch welche Merkmale jeder einzelne Cluster am stärksten gekennzeichnet ist (Ääremaa, 1991). Somit stellt dieses Programmsystem eine Kombination von der Diskriminanz- und Clusteranalyse sowie von einigen Elementen der Faktorenanalyse dar (Hartung / Elpelt, 1986) und ist u.E. besonders geeignet für die Analyse eines schwer überschaubaren Belegmaterials.

2. Resultate der Untersuchung einer Gruppe deutscher Aussehensadjektive

Die oben beschriebenen Methoden wurden bei der Analyse einer Gruppe deutscher Adjektive angewendet. 34 einfache Lexeme, die bei der Beschreibung des menschlichen Aussehens (attributiv) gebraucht werden und die semantische Komponente einer ästhetischen Wertung enthalten, wurden auf ihre denotativen Merkmale hin untersucht, um die semantische Struktur der Gruppe festzustellen. Die Merkmale wurden wie folgt formuliert:

+BEWER 'positive Bewertung'
- BEWER 'negative Bewertung'
+DIM 'große Dimension'
- DIM 'kleine Dimension'
KONST 'Konstitution'
FORM 'Form'
GEORD 'Geordnetheit'
REIN 'Reinheit'
KLEID 'Kleidung'
BELEB 'gesundheitlicher Zustand'
+BEWEG 'Bewegung'
PSYCH 'psychische Charakteristika (Stimmung, Verhalten bzw. Charakter)'
REIZ 'Reiz, Eindruck'

Die Markiertheit eines Objektes durch ein bestimmtes Merkmal wird durch den Merkmalswert "1" bezeichnet, die Nichtmarkiertheit durch "0". Nach diesem Prinzip wurde eine Datenmatrix zusammengestellt (s.Tab.1).

Bei so allgemein formulierten Merkmalen ist es unvermeidlich, daß die Merkmalkombinationen mancher Objektpaare (bzw. -gruppen) völlig übereinstimmen. Gegebenenfalls sind

Tabelle 1

Merkmalswerte ästhetisch bewertender Aussehensadjektive

Merkmal / Objekt	+ BEWER	- BEWER	+ DIM	- DIM	KONST	FORM	GEORD	REIN	KLEID	BELEB	+BEWEG	PSYCH	REIZ
edel	1	0	0	0	0	1	1	0	0	0	0	0	1
elend	0	1	0	0	0	0	0	0	0	1	0	0	1
fein$_1$	1	0	0	1	1	1	0	0	0	0	0	0	0
fein$_2$	1	0	0	0	0	0	0	1	0	0	0	0	0
fein$_3$	1	0	0	0	0	0	0	0	1	0	0	1	0
feist	0	1	1	0	1	0	0	0	0	0	0	0	0
fesch	1	0	0	0	0	0	0	0	1	0	1	1	0
flott	1	0	0	0	0	0	0	0	0	0	1	1	1
frisch	1	0	0	0	0	0	0	0	0	1	0	0	1
grob	0	1	0	0	1	1	0	0	0	0	0	0	1
hehr	1	0	0	0	0	0	0	0	0	0	0	1	1
herb	0	1	0	0	0	0	0	0	0	0	0	1	1
hold	1	0	0	0	0	0	0	0	0	0	0	0	1
hübsch	1	0	0	0	0	0	0	0	0	0	0	0	1
keck	1	0	0	0	0	0	0	0	0	0	0	1	1
keß	1	0	0	0	0	0	0	0	1	0	0	1	1
lecker	1	0	0	0	0	0	1	1	1	0	0	0	1
nett	1	0	0	0	0	0	1	1	1	0	0	0	1
nobel	1	0	0	0	0	0	0	0	1	0	0	1	0
pikant	1	0	0	0	0	0	0	0	0	0	0	0	1
plump	0	1	1	1	1	1	0	0	0	0	0	0	0
proper	1	0	0	0	0	0	1	1	1	0	0	0	0
rank	1	0	1	1	1	1	0	0	0	0	1	0	0
resch	1	0	0	0	0	0	0	0	0	0	0	1	1
sauber	1	0	0	0	0	0	0	0	0	0	0	0	1
schau	1	0	0	0	0	0	0	0	0	0	0	0	0
schick	1	0	0	0	0	0	1	0	1	0	0	0	0
schlank	1	0	1	1	1	0	1	0	0	0	0	0	1
schmuck	1	0	0	0	0	0	0	0	1	0	0	0	1
schnieke	1	0	0	0	0	0	0	0	1	0	0	0	1
schön	1	0	0	0	0	0	0	0	0	0	0	0	1
smart	1	0	0	0	0	0	0	0	1	0	1	1	0
süß	1	0	0	0	0	0	0	0	0	0	0	0	1
zart	1	0	0	1	1	0	0	0	0	0	0	0	1

das die Gruppen *lecker - nett - proper, hehr - keck - resch* und *hold - hübsch - pikant - sauber - schön - süß* und die Lexempaare *schmuck- schnieke, fesch - smart* und *fein$_2$ - nobel*. Die gegenseitige Distanz zwischen den Elementen jeder Gruppe ist hier gleich 0 ($d = 0$). Es handelt sich aber keineswegs um vollständige Synonyme, denn viele Merkmale, wie z.B. stilistische bzw. territoriale Besonderheiten, mögliche Kollokationspartner sowie die feinere Strukturierung der Merkmale sind nicht berücksichtigt worden. Es ist aber auch nicht zweckmäßig, alle möglichen Merkmale gleichzeitig zu behandeln, weil es sich um Merkmale verschiedenen Grades handelt, und ihr Behandeln als gleichwertige Merkmale das Gesamtbild entstellen könnte. Es ist wohl möglich, das mittels der Analyse der Merkmale einer Ebene ermittelte Clustersystem mit einem anderen zu vergleichen, in dem die Merkmale einer anderen Ebene in die Analyse mit einbezogen sind. Auf diese Weise kann man feststellen, wie die neu einbezogenen Merkmale das Anfangsbild ändern. Als Grundlage müßten aber unbedingt die Merkmale ein und desselben Grades dienen.

2.1. 1-Clusterung

Anschließend beschreiben wir den Verlauf der 1-Clusterung, wobei wir die durch verschiedene Methoden gewonnenen Resultate miteinander vergleichen (s. Tab.2).

Wie oben gesagt, vereinigen sich auf dem ersten Schritt der Clusterung ($h = 0$) die Adjektive, deren Abstand gleich 0 ist:
C1 {lecker, nett, proper};
C2 {fein$_2$, nobel};
C3 {hehr, keck, resch};
C4 {schmuck, schnieke};
C5 {fesch, smart};
C6 {hold, hübsch, pikant, sauber, schön, süß}.

Der Vergleich des weiteren Verlaufs der Clusterung aufgrund verschiedener Distanzformeln hat ergeben, daß sich die Objekte mit wenigen durch den Merkmalswert "1" verzeichneten Merkmalen (bei dem Binärcode) aufgrund der euklidischen Distanz (d_E) früher dem Zentrum anschließen als aufgrund der linearen Korrelation und des Sörensen-Ähnlichkeitskoeffizienten. Diese Tatsache ist dadurch zu erklären, daß bei der d_E alle Merkmalswerte berücksichtigt werden, d.h., die Nichtmarkiertheit eines Objekts durch ein bestimmtes Merkmal gilt ebenso als ein Merkmal (nur mit einem "0"-Wert). Bei der Clusterung aufgrund der Distanz d_L sowie der d_S wird aber der Anzahl der durch "1" verzeichneten Merkmale bei einem Objekt eine größere Bedeutung beigemessen. Je größer diese Zahl ist, desto größer ist auch die potentielle Zahl der mit anderen Objekten gemeinsamen Merkmale und dadurch die Möglichkeit, mit mehreren Objekten verbunden und zum Zentrum gerechnet zu werden. Diese statistischen Nuancen ändern aber in unserem Falle nicht das Endergebnis, das in folgenden Postulaten zusammengefaßt werden kann:

1) Die Kernzone der ästhetisch bewertenden einfachen Aussehens-adjektive umfaßt
- eine allgemeine positive Bewertung tragende Adjektive *(hold, hübsch, pikant, sauber, schön, süß, schau, fein$_3$, edel, frisch)* (die drei letztgenannten enthalten neben der allgemeinen Bewertung auch einen Verweis auf die Reinheit (Farbe), Form bzw. Gesundheit);
- kleidungsbezogene Adjektive *(schmuck, schnieke, keß, lecker, nett, proper, fein$_2$, fesch, nobel, smart,*

Tabelle 2

Verlauf der 1-Clusterung aufgrund verschiedener Distanzkoeffizienten

d_S	d_L	d_E	d_{HC}
C1 { lecker, nett, proper }			
C2 { fein$_2$, nobel }			
C3 { hehr, keck, resch }			
C4 { schmuck, schnieke }			
C5 { fesch, smart }			
C6 { hold, hübsch, pikant, sauber, schön, süß }			
C7 { C2, C3, C4, C5; flott, keß }	C7 {C2,C3,C4,C5,C6; fein$_3$, flott, frisch, keß, schick}		C7 { C3; flott }
C8 { C6, C7; frisch }			C8 { C6; frisch }
C9 { fein$_1$, rank }			C9 { C4; keß }
C10 { schlank, zart }			C10 {C10;fein$_3$,schau}
			C11 { C2, C5 }
			C12 { C8, C9 }
			C13 { C7, C12 }
C11 { feist, plump }	C8 {feist, plump}		C14 { C1; schick }
C12 { C1, C8; schick }	C9 { C1, C7; edel, elend, fein$_1$,herb, rank, schick, schlank, zart }		C15 { schlank, zart }
C13 { C9, C10 }	C13{C12; fein$_3$,schau}		C16 { fein$_1$, rank }
			C17 { feist, plump }
C14 { C11, C13 }	C14 {C10, C13; edel}		C18 { elend, herb }
	C15 {C9, C14 }		C19 {C13, C10 }
			C20 {C19; edel }
	C16 {C11, C15 }		C21 {C14, C20 }
			C22 {C11, C21 }
C15 { C12, C14; edel, elend, fein$_3$, grob, herb, schau}	C17 {C16; elend,herb }	C10 {C8, C9; grob }	C23 {C15, C16 }
			C24 {C18; grob }
	C18 {C17; grob }		C25 {C17, C22 }
			C26 {C24, C21 }
			C27 {C26, C25 }

schick);
- psychikbezogene Adjektive *(fein2, fesch, flott, hehr, keck, keß, nobel, resch, smart)*.

2) Zur Peripherie gehören
- Dimensions-Konstitutionsadjektive *(schlank, zart, rank, fein1, feist, plump)*,
- eine negative Bewertung tragende Adjektive *(elend, grob, herb, feist, plump)*.

3) Im allgemeinen vereinigen sich zuerst die Wörter mit ähnlicher Bedeutung (Synonyme) (z.B. {fein1, rank}, {schlank, zart}, {plump, feist}. Auf dem nächsten Schritt schließen sich ihnen die Wörter mit der entgegengesetzten Bedeutung (Antonyme) an (bei d_S: C14 {feist, plump, fein1, rank, schlank, zart}; ebenso bilden sich bei d_L der Cluster C16, bei d_E der Cluster C10, bei d_{HC} die Clusters C25, C27).

2.2. 2-Clusterung

Als Nachteil der 1-Clusterung gilt, wie oben gesagt, die Tatsache, daß der gegenseitige Abstand der Objekte, die zu einem Cluster zusammengeschlossen werden, hier nicht zum Ausdruck kommt (einigermaßen wird die feinere Struktur der Clusters nur durch die Zentroid-Methode sichtbar). Demgegenüber ist es bei der 2-Clusterung möglich, daß ein Objekt gleichzeitig zu mehreren Clusters gehören und auf diese Weise als verbindendes Glied zwischen den voneinander weiter entfernt liegenden Objekten dienen kann. Die Resultate der Analyse lassen sich graphisch darstellen (s. Abb.1).

Aufgrund der 2-Clusterung kann man folgendes feststellen:

1) Den Kern der behandelten Adjektivgruppe bilden die Adjektive, die einen angenehmen Eindruck (Reiz) auf die Sinnesorgane eines Menschen ausdrücken, der durch die äußere Erscheinung eines Menschen (bzw. seiner Körperteile), gelegentlich auch durch seine Kleidung und / oder sein Verhalten (bzw. Stimmung, Charakter) hervorgerufen ist. Zu dieser Gruppe gehören die Adjektive: *hold, hübsch, pikant, sauber, schön, süß, frisch, schmuck, schmieke, resch, hehr, keck, flott, keß*.

2) Mit der Kerngruppe sind eng verbunden die Adjektive, in denen die positive Wertung nicht so sehr auf einem Sinneseindruck, als vor allem auf einer rational begründeten Einschätzung beruht:

- auf die Psychik bezogene Adjektive: *fesch, smart, fein2, nobel;*

- auf die Kleidung bezogene Adjektive: *lecker, nett, proper*.

3) Zur Randzone der zentralen Gruppe gehören die Adjektive *edel, schau, herb, zart, schick*, die die zentrale Gruppe gelegentlich mit anderen (peripheren) Gruppen verbinden.

4) Über das Adjektiv *zart* schließt sich der zentralen Gruppe die größte periphere Gruppe an, die ihrerseits in folgende Untergruppen eingeteilt ist:
- eine positive Bewertung enthaltende Dimensions-Konstitutionsadjektive (*zart, schlank, fein1, rank);*
- Konstitutions-Formadjektive *(rank, fein1, plump);*
- eine negative Bewertung tragende Dimensions-Konstitutionsadjektive *(plump, feist)*;
- eine negative Bewertung tragende

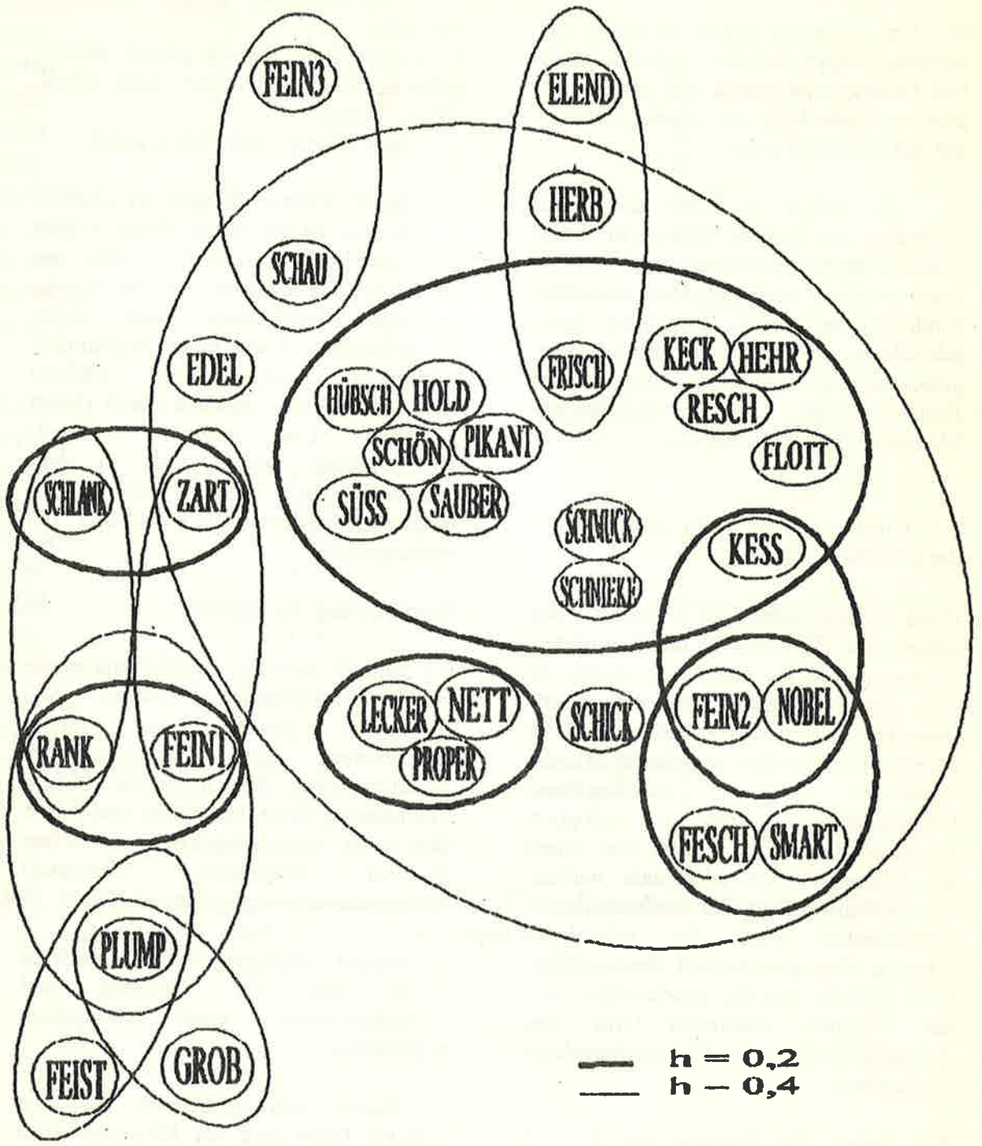


Abb.1. 2-Clusterung der einfachen Aussehensadjektive

Konstitutions-Formadjektive *(plump, grob)*.

5) Das Adjektiv *schau* verbindet die zentrale Gruppe mit dem Adjektiv *fein*$_3$ (die Untergruppe enthält eine allgemeine positive Bewertung mit einem Verweis auf die Reinheit/Farbe).

6) Die Adjektive *frisch* und *herb* verbinden die zentrale Gruppe mit dem Adjektiv *elend* (die Untergruppe enthält einen positiven oder negativen visuellen Eindruck, der durch den guten bzw. schlechten Gesundheitszustand oder gelegentlich auch durch den psychischen Zustand des zu beschreibenden Menschen hervorgerufen ist).

2.3. Clustering mit Hilfe der Methode der *k*-Mittelpunkte

Beim Interpretieren der Resultate der mittels der B_k-Methode durchgeführten Clusteranalyse ist man gezwungen, zu den Ausgangsdaten zurückzugreifen, denn die Distanzmatrizen enthalten keine Information darüber, welche konkreten Merkmale bei verschiedenen Objektpaaren bzw. -gruppen übereinstimmen und welche von ihnen als differenzierende Merkmale wirken. Bei der Anwendung der Methode der *k*-Mittelpunkte wird die inhaltliche Analyse aber automatisch durchgeführt. Man braucht nur die gewünschte Zahl der Clusters zusammen mit den Ausgangsdaten in den Elektronenrechner einzugeben.

Wir haben die Analyse bei $k = 5$ durchgeführt, d.h., es wurde im voraus bestimmt, daß wir die uns interessierenden Adjektive in fünf Clusters einteilen wollen. Der Elektronenrechner hat dei Clusters wie folgt gebildet:

C1 {flott, hehr, herb, keck, keß, resch};
C2 {fein$_2$, fesch, nobel, schick, smart};
C3 {lecker, nett, proper, schmuck, schnieke};
C4 {frisch, hold, hübsch, pikant, sauber, schlank, schön, süß, zart, edel, elend, schau, fein$_3$};
C5 {feist, plump, rank, fein$_1$, grob}.

Von den Profildarstellungen der Clusters (s. Abb. 2) ist zu sehen, durch welche Merkmale jeder einzelne Cluster am stärksten gekennzeichnet ist. Die von der zentralen Vertikallinie nach rechts abweichenden Werte eines bestimmten Merkmals bezeichnen die stärkere Markiertheit der Adjektive durch dieses Merkmal. Die großen Standardabweichungen zeigen, daß in dem entsprechenden Cluster die beiden möglichen Merkmalswerte ("1" und "0") vertreten sind.

Beschreibung der Clusters:

C1 enthält sowohl positiv als auch negativ bewertende Adjektive. Die Bewertung beruht auf einem angenehmen bzw. unangenehmen visuellen Reiz, der durch die äußere Erscheinung eines Menschen und durch den darin widerspiegelten psychischen Zustand (Verhalten, Charakter) hervorgerufen ist.

C2 enthält Adjektive, die ein positives Urteil über die Kleidung und Verhaltensweise eines Menschen ausdrücken.

C3 besteht aus Adjektiven, die eine positive Bewertung der Kleidung eines Menschen, insbesondere deren Reinheit, Geordnetheit und den dadurch hervorgerufenen angenehmen Eindruck ausdrücken.

C4 umfaßt Adjektive mit einem

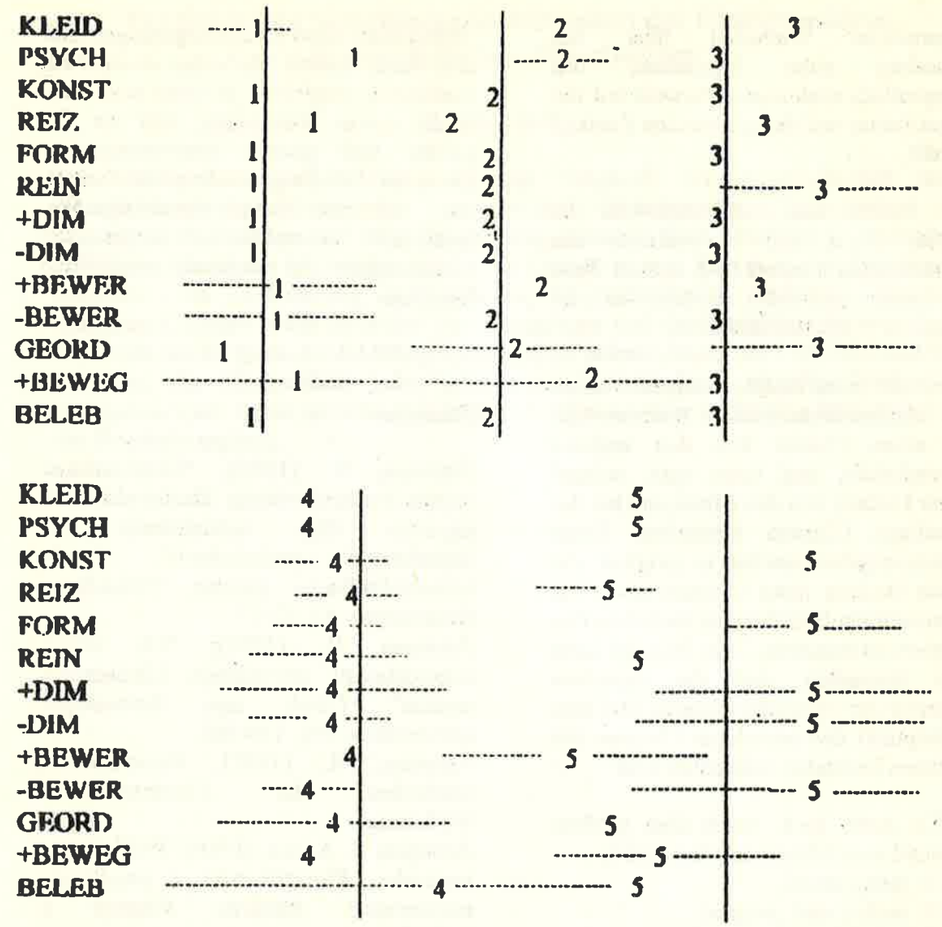


Abb.2. Profildarstellungen der Clusters C1 - C5 aufgrund der Clusteranalyse mit Hilfe der Methode der k-Mittelpunkte

allgemeinen Werturteil über das Aussehen eines Menschen, das gelegentlich auch einen Verweis auf die Konstitution und den physischen Zustand enthält.

C5 besteht aus Adjektiven, die die Größe, Form und Konstitution des menschlichen Körpers bzw. dessen Teile beschreiben und dabei ein positives oder negatives Werturteil enthalten.

Inwieweit jedes einzelne Adjektiv durch die Merkmalkombination markiert ist, die einen Cluster von den anderen unterscheidet, das kann man anhand seiner Distanz von den Mittelpunkten der jeweiligen Clusters feststellen. Diese Distanzangaben machen es möglich, die innere Struktur jedes Clusters sowie die gegenseitigen Beziehungen zwischen den Clusters zu ermitteln. Zum Beispiel kann man feststellen, daß die typischen Vertreter der einzelnen Clusters (die dem Mittelpunkt des jeweiligen Clusters am nächsten liegenden Adjektive) sind:

in C1: *hehr, keck, resch* (den größten Abstand vom Mittelpunkt hat *herb*);
in C2: *fein2, nobel*;
in C3: *lecker, nett, proper*;
in C4: *hold, hübsch, pikant, sauber, schön, süß* (den größten Abstand vom Mittelpunkt haben *schlank* und *elend*);
in C5: *plump* (den größten Abstand haben *grob* und *rank*)

Von der Distanzmatrix der Mittelpunkte der Clusters ist auch abzulesen, daß der Cluster C5 von allen anderen Clusters mehr entfernt ist als die anderen Clusters voneinander. Am nächsten liegen zueinander die Clusters C1 und C4. Auch die Clusters C1 und C2, C3 und C4 sowie C2 und C3 sind miteinander ziemlich eng verbunden.

Wenn man diese Schlußfolgerungen mit den durch andere Methoden ermittelten Resultaten vergleicht, so kann man sich leicht davon überzeugen, daß sie im großen und ganzen übereinstimmen. Dabei ist aber die gute Interpretierbarkeit der Methode der *k*-Mittelpunkte besonders hervorzuheben wegen der Vielseitigkeit der dadurch ermittelten Resultate.

Literatur:

Ääremaa, R. (1978), "Obschtschaja teorija konstruirovanija klaster-sistem i algoritm dlja nahozhdenija ih tschislennyh predstavlenij", *Trudy vytschislitel'nogo tsentra Tartuskogo universiteta,* 42, 53-77.

Ääremaa, R. (1981), "Ob odnoj vozmozhnosti provedenija klasternogo analiza", *Utsch. zap. Tartuskogo universiteta,* 591, 158-162.

Ääremaa, R. (1991), *Verschiedene Methoden der Clusteranalyse.* Vorlesung.

Aivazjan, S. A. u.a. (1989), *Prikladnaja statistika. Klassifikatsija i snizhenije razmernosti.* Moskva: Finansy i statistika.

Hartung,J./Elpelt, B. (1986), *Multivariate Statistik. Lehr- und Handbuch der angewandten Statistik.* 2., überarbeitete und ergänzte Auflage. München - Wien: R. Oldenbourg.

Tuldava, J. (1987), *Problemy i metody kvantitativno- sistemnogo issledovanija leksiki.* Tallinn: Valgus.

Vergleich von verschiedenen Methoden der Clusteranalyse in linguistischen Forschungen

Resümee

Die Untersuchung befaßt sich mit dem Vergleich der Anwendbarkeit und Interpretationsmöglichkeiten von verschiedenen Varianten der Clusteranalyse in linguistischen Forschungen. Als Untersuchungsmaterial dient eine Gruppe deutscher Adjektive. Durch die 1-Clusterung wird die Gruppe in eine Kernzone und Peripherie aufgeteilt. Die 2-Clusterung ermöglicht es, die verbindenden Elemente zwischen einzelnen Clusters sowie zwischen der Kerngruppe und der Peripherie festzustellen. Mit Hilfe der Methode der *k*-Mittelpunkte wird zu jedem Cluster ein Profilbild erstellt, auf dem die Markiertheit jedes gebildeten Clusters durch einzelne Merkmale angegeben wird. Somit wird die quantitative Analyse um eine qualitative Beschreibung der Resultate ergänzt.

Статистический анализ лексического состаа текстов функциональных стилей.

Баскевич Валентина Михайловна
Винницкий Государственный Технический Университет
Кафедра иностранных языков
286037 Винница / Украина, ул. Келецкая 39, кв.193
тел. (0432) 43 49 88

Доклад.

ТЕМАТИЧЕСКАЯ ОБЛАСТЬ: Статистический анализ текста; функциональная стилистика

Резюме:
Исследовано статистическое поведение семантических подклассов слов в текстах газетно-публицистического стиля и художественной прозы. В результате теста на "стилевую маркированность" подкласса с помощью критерия ХИ-квадрат выделены подклассы слов, характерные для лексического наполнения текстов газет и художественной прозы. Тест на "статистическую однородность" показал регулярное соотношение частот подклассов в подвыборках из текстов разных газет, что свидетельствует об "организменности", т.е. согласованности лексических структур газетных текстов.

Лексику текстов функциональных стилей сравнивают на двух уровнях: на уровне словаря, когда исходят из набора слов (лексем) без учета их частотности, или на уровне текста, когда фиксируют частоты слов в конкретных высказываниях (текстах). При этом, как правило, проводится сопоставление частотных характеристик отдельных слов. Богатый опыт системного описания лексики делает целесообразным и необходимым изучать поведение также лексико-семантических разрядов или подклассов слов в текстах разных функциональных стилей. Статистические параметры семантических подклассов слов как формальные свойства лексико-семантической организации текстов позволят измерить лексическую связь текстов и раскрыть качественные особенности стилей.

В настоящей работе сравниваются на уровне текста количественные параметры семантических подклассов прилагательных, функционирующих в текстах газет и художественной прозы современного немецкого языка.

Методом сплошной выборки зафиксировано 16 тыс. словоупотреблений прилагательных в атрибутивной функции в текстах немецких газет и 4 тыс. словоупотреблений прилагательных в текс-тах художественной прозы. Выделено 19 семантических подклассов прилагательных, которые обозначают, например:
- общественно-политические отношения (politisch, interna-
tional, demokratisch, gesellschaftlich);
- периоды времени и возраст (gegenwaertig, dreitaegig, jung);
- положительную оценку (positiv, gut, praechtig);
- интенсивность (fest, scharf, schwer, beharrlich);
- динамичность (aktiv, dynamisch, rasch, schnell);
- локальные характеристики (europaeisch, westlich, oert-
lich);
- свето-цветовые признаки (dunkel, schwarz, rot, gruen);
- температуру (warm, heiss , kalt и др.) см. табл.1).
Показатели частот подклассов в текстах исследуемых стилей
приведены в исходной табл.1.

1.Тест на "стилевую маркированность" подкласса.

Под "стилевой маркированностью" будем понимать закрепленность лексики за определенными стилевыми разновидностями языка. Применительно к настоящему исследованию данное понятие предполагает повышение фактически наблюдаемых частот подклассов прилагательных над теоретически ожидаемыми величинами в текстах того или иного стиля.

Таблица 1

Распределение частот подклассов прилагательных в текстах двух стилей

Подклассы N. прилагательных	Выборки	
	Газетн.-публиц.	Худ. проза
1 Общественно-политические отношения	2452	104
2. Размер	2354	697
3. Положительная оценка	2166	409
4. Периоды времени и возраст	1574	537
5. Пространственные характеристики	1255	319
6. Термины	1165	29
7. Интенсивность	1031	330
8. Внутреннее состояние	811	233
9. Локальные характеристики	840	142
10 Экономические отношения	589	10
11 Динамичность	555	63
12 Интеллект	537	119
13 Отрицательная оценка	192	175
14 Свето-цветовые признаки	94	288
15 Внешние признаки	100	248
16 Температура	32	31
17 Физическое и психическое состоя-ние человека	37	51
18 Материал	18	41
19 Другие признаки и свойства	198	174
Всего:	16000	4000

С этой целью с помощью критерия ХИ-квадрат (X^2) устанавливались статистически значимые расхождения между эмпирически наблюдаемыми и теоретически ожидаемыми частотами подклассов прилагательных в текстах газет и художественной прозы (см. табл.2).

Таблица 2

Альтернативная таблица для определения статистически значимого расхождения между подклассами в двух стилях

Подкласс	Газетн.-публиц.	Худ. проза	Всего
Общ.-полит. отношения	2452	104	2556
Другие подклассы	13548	3896	17444
Всего	16000	4000	20000

Вычисленный на основе табл.2 $X^2 = 464,85$. Статистически значимыми считались расхождения при $X^2 \geq 3,84$.

Результаты анализа показали, что "маркированными", т.е. характерными для текстов газетно-публицистического стиля, оказались, как и следовало ожидать, семантические подклассы прила-гательных, обозначающих:
- общ.-политические отношения ($X^2 = 464,85$)
- положительную оценку 31,29
- термины 244,73
- экономические отношения 136,64
- интеллект 4,14
- динамичность 40,26
- локальные характеристики 23,55

Для текстов художественной прозы "маркированными" оказались подклассы прилагательных, которые обозначают:
- размер ($X^2 = 18,2$)
- периоды времени и возраст 43,61
- интенсивность 16,44
- отрицательную оценку 40,26
- свето-цветовые признаки 746,82
- внешние признаки 581,69
- температуру 33,68
- физическое и психическое состояние человека 79,57
- материал 92,49

Для подклассов прилагательных с семантикой пространственных характеристик и внутреннего состояния не выявлены статистически значимые различия в двух стилях. Очевидно, прилагательные данных подклассов занимают сравнительно одинаковое положение в ранжированных частотных рядах исследуемых стилей.

Разное положение подклассов прилагательных в частотных ранжированных рядах двух стилей продиктовано, главным образом, экстралингвистическими факторами. Предполагается, что ярко выраженная референтная отнесенность текстов газет к сфере общественно-политической жизни определяет денотативный и сигнификативный аспекты их содержания и, тем самым, набор основных семантических подклассов прилагательных, характерных для лексико-семантического наполнения текс-
тов данного стиля.

2. Тест на статистическую однородность

Как известно, данные о закономерности языковых явлений можно получить лишь тогда, когда явления носят регулярный характер. Понятие регулярности коррелирует с такими понятиями, как устойчивость статистических рядов, нормативность (Пиотровский Р.Г., Бектаев К.Б., 1977; Тулдава Ю.А., 1987), стабильность (Андрющенко В.М., 1978), организменность (Арапов М.В., 1982; 1988), которые используются авторами в качестве критериев статистической однородности. Общее понимание статис-тической однородности, как оно определено в статистике и ис пользуется в лингвистических исследованиях, - "...статистическая однородность вытекает из понятия генеральной совокупности, т.е. такой наиболее общей совокупности фактов, в которой они существенно не отличаются по своим статистическим характеристикам" (Перебейнос В.И., Муравицкая М.П., Дарчук Н.П., 1985), детализируется в данной работе и рассматривается уже как результат соблюдения требований к лингвистической однородности. Такой план исследования подтверждает истинность эмпирических наблюдений, способствуя объективному выявлению закономерностей функционирования языковых единиц.

Применительно к данному исследованию тест на статистическую однородность предполагает выявление характера частотного распределения (положения) семантических подклассов прилагательных в разных подвыборках газет, а также в выборках, принадлежащих к разным стилям. В качестве оперативного приема проверки критериев однородности использован метод корреляции рангов по Спир-мену. Частоты подклассов прилагательных в 10-ти подвыборках были зафиксированы в исходных таблицах. Вычислительные операции проводились с помощью ЭВМ. Проведено 20 сопоставлений подвыборок друг с другом. Величины коэффициентов ранговой корреляции (ρ) представлены в таблице 3.

Высокие значения коэффициентов для д подвыборок из текстов газетно-публистического стиля, превышающие $\rho_{0,05} = 0,64$ показали, что относительное положение подклассов прилагательных в ранжированном ряду для подвыборок из разных газет приблизительно одинаковое.

Таблица 3

Результаты исследования ранговой корреляции между подвыборками

Подвы-борки	Газетно-публицистический стиль					Худ. проза				
	1	2	3	4	5	6	7	8	9	10
1	x	0,99	0,97	0,87	0,86	0,96	0,94	0,95	0,44	0,38
2.		x	0,97	0,98	0,96	0,96	0,91	0,94	0,56	0,38
3.			x	0,97	0,96	0,96	0,91	0,93	0,64	0,43
4.				x	0,98	0,98	0,90	0,92	0,62	0,36
5.					x	0,99	0,90	0,93	0,61	0,33
6.						x	0,91	0,93	0,61	0,33
7.							x	0,98	0,49	0,34
8.								x	0,53	0,36
9.									x	0,79
10.										x

Такое регулярное соотношение частот подклассов прилагательных в подвыборках из

текстов разных газет свидетельствует об "организменности" /Арапов М.В., 1982 /, которая представляет собой согласованность лексических структур всех текстов одного стиля.

Если исходить из того, что каждый семантический подкласс имеет свою тему, то результаты анализа говорят в пользу "стереотипности содержания" в газете: количество тем, по которым ежедневно высказывается газета, поддается учету - подкрепляется стереотипностью языкового оформления.

Полученный коэффициент ранговой корреляции для подвыборок из текстов художественной прозы (p = 0.79) несколько меньше, чем величина p для подвыборок из текстов газет, хотя и превышает критическое значение табличного коэффициента. Следовательно, в текстах художественной прозы действуют менее строгие правила лексического наполнения текстов.

Сравнение величин p для подвыборок из разных стилей, а они показывают существенное расхождение (см. подчеркнутые коэффициенты в табл. 3), подтверждает тот факт, что закономерности употребления семантических подклассов прилагательных в двух стилях носят разный характер. Употребление прилагательных тех или иных подклассов, их частота обусловлены целями и задачами функционального стиля.

Проведенное исследование убеждает нас в том, что газетно-публицистический стиль обладает выдержанным типичным единством, проявляющимся в устойчивой организации языкового материала в связи с целями функционального использования.

ПРИМЕЧАНИЯ

1) Корреляция присутствовала, если полученные $p > p_{табл}$, т.е. $p>0{,}71$ при n=10; корелляция отсутствовала, если $p<p$, т.е. $p<0{,}64$ при n=10, где n - количество подвыборок.

Statistical analysis of lexical text composition of functional styles

Valentina M.Baskevich

Summary:
The statistical behaviour of the semantic word subclasses in the newspaper and fiction text styles has been investigated. As a result of "style marking" test by means of XI-square criterion the word subclasses characteristic of newspaper and fiction text vocabulary have been distinguished. The test for "statistical homogeneity" made with the help of Spirman correlation method showed a regular frequency correlation of the word subclasses in the sample texts of various newspaper. It proves the coordination of lexical text tructures of the newspaper-styles

Theoretical Principles of Linguistic Database Design in Natural Language Processing

Larissa N. Beliaeva,
State Pedagogical University of Russia,
Mashine Translation Laboratory,
Russia,191183, Saint Petersburg,Moika 48,

Topical paper

AREA: Computational Linguistics, Automatic Translation

Summary:
In the paper there are discussed quantitative, linguistic and computer problems of designing databases for natural language processing systems on the example of databases for machine translation, which are realized in SILOD system.

Databases which are designed for various intellectual systems differ about their structure, composition, type of components, set of information and relation systems between the elements.

But in spite of their differences all possible databases of expert systems have common features and common problems which are to be solved when designing a database. These problems are concerned with linguistic nature of elements with the help of which any database item can be described. As natural language words are the only feasible facilities to define concepts or items of any human activity, language processors and special linguistic databases are integral parts of any serious expert system.

Thus when creating a database the designer (linguist, programmer or knowledge engineer) faces linguistic bottlenecks which are common for all natural language processing (NLP) systems. As the main field of our NLP investigations is Machine Translation (MT) which in some cases can be considered as a part of expert system, we'll discuss linguistic problems in question on the example of linguistic databases for MT. Such databases can be realized as a system of automatic dictionaries (AD) or as one integral AD.

It is to be emphasized, that the main difference between AD for MT and databases for expert systems of various complexity is the mode of information representation. It means that in AD all kinds of information are oriented on computer processing and therefore are to be represented as special codes. The details of such coding depend on the peculiarities and degree of development of a specific NLP system.

In case of expert system databases the mode of information representation must be oriented both on computer processing and knowledge defenition for a user.

1. Automatic Dictionary Items in Linguistic Databases

Dependency of the database structure on the knowledge domain and main task of a NLP system and as a consequence, the necessity of AD adjusting to the domain peculiarities is now mutually recognized. The same refers to the volume of a NLP-system database. It is now absolutely clear that creation of a practically usable expert system requires to design a huge database, items of which can represent the main concepts and conventional terminology of the domain in question. Not less than 95% of the source text items are to be distinguished and described with the help of a database if the expert or NLP system is oriented for a practical use.

Naturally, particular volume of a database depends on the typology of the source language and the chosen procedure of morphological analysis, the aim of which is speedy and accurate identification of the source text wordforms with the help of AD.

The choice of a particular form of AD is dependent on two main parameters:

- time-consuming parameter, which shows the time necessary for a text wordform identification;
- stability of morphological procedures which are to ensure word-forms identification and prevent improper identification.

Thus, we can see that the first task of AD in any NLP or expert system is text word-form identification, procedures of which depend on the chosen type of machine morphology and as a consequence on the type of AD items. The choice of AD item is determined both by word- and formbuilding principles different in specific languages as well as by the representation features of semantic items of a text.

Besides, the choice of a basic dictionary item is determined by the tasks of NLP system. Our experience in Designing AD for typologically-different languages has shown that for analytical languages AD created as a set of separate word-forms is the most expedient mode, as it allows to increase the speed of the system while the growth of the dictionary volume is negligible. For synthetical languages adoption of special computer methods of morphological analysis is equally expedient. In this case machine stems are considered as the heads od dictionary items (DI) and AD is to be provided with machine morphology.

The concept of the machine affix separation had allowed to elaborate principles of machine morphology creation, universal for many flective and agglunative languages (for example, now these principles are justified for all Roman languages, Russian and Greek languages and for some part of Finnish language morphology). Machine morphology is formed as a set of paradigms - machine affix chains. Each typical paradigm correlates with the grammatical characteristics of the stems and the word formation mode. The link between a machine stem and a paradigm is realized with the help of a special code, which characterizes all possible word-forms which can be generated from the stem in question .

The result of morphological analysis, which is received with the help of AD and special lexical and morphological analysis algorithms, is a source for NLP algorithms.

Irrespective of the use of machine morphology any AD includes both word-forms and stems, because always there are some cases which make usage of word-forms advantageous (for example, it is expedient to include all forms of modal and auxiliary verbs in AD of any language).

Requirement to include into AD both separate lexical units as word-forms and stems and combinations of such elements (machine phrases) is recognized now by all database designers. But the bottleneck of such automatic machine phrases dictionaries (AMPD) lies in the necessity to establish for any database the following:

- typology of machine phrases;
- method of their recognition in course of text analysis;
- method of AMPD storing.

The problem of AMPD is connected with the fact that new and important notions, in all contemporary languages, are often expressed by means of phrases. Therefore when creating a linguistic database it is essential to elaborate a frequency dictionary (FD) of phrases for every application domain alongside with the traditional FD of words. Such domain-oriented FD of phrases can be used to compile huge AD of phrases and words and to solve the problems which were listed above.

From typological point of view it is possible to determine the following kinds of machine phrases (MP):

- icon MP, i.e. unchangeable linear combinations of wordforms, functional, semantic and syntactic characteristics of which do not depend on the context (in all languages we can find composite prepositions, conjunctions, adjuncts etc. as such MP);
- icon changeable MP, which are represented by linearly continuous sequences of words, functional characteristics of which depend on the syntactic function of a MP in the sentence. As an example of such MP we can present terminology (see terms 'high-level rack storage' (Eng.) - 'etagere multiictages' (Fr.) - 'многоярусный стеллаж' (Rus.) in which the structure of the terms do not coincide when the meaning is the same);
- conventional iconical MP, which are represented by unchangeable sequences of words, sunctional and semantic characteristics of which depend on the context, in particular on presence of a punctional mark, which can isolate a parenthetic clause: for example, 'as a whole,' (Eng.), the same for 'в целом,' (Rus.), but it is possible to use this construction in other structures: 'as a whole series' and 'как целый ряд';
- discontinuous MP, which are represented as sequences of words, which are integral unities from semantic and functional points of view but can be separated with other words or phrases. In the languages with free order of words the components of such MP can take up any text position in relations to the kernel word. As an example of such MP there can be used verbal phrases. As to the morphology the discontinuous MP can be both changeable and unchangeable.

Naturally, in the course of lexical-and-morphologocal text analysis only proper icon MP and icon changeable MP can be identified, the two last types of MP are to be specially analyzed and identified on different parsing levels.

2. Dictionary Item Structure

Thus, any linguistic database, which can be a part of MT system or a special entry to knowledge or terminology database of any expert system, at least includes source word dictionaries, which are organized both as dictionaries of word usages and dictionaries of stems, source phrase dictionaries and machine morphology for different languages.

AD of any NLP system is its nucleus, because the lingu- and software can be realized just on the basis of the information which is stored in the AD. So a special consideration must be given to the volume of information which is to be ascribed to any AD element and to the mode of its storing in the AD and extracting from it.

Experience of practical MT system designing had shown that it is impossible to elaborate a complete structure of DI for a NLP system ad hoc, at once and for all theoretially probable situations. Even if the procedure of creating a word portrait is attractive for a linguist, in reality we must include in the DI only the information which is justified by realized algorithms.

Naturally, such approach must be added with creation of special procedures and conditions that allow to complement any DI with new information which is acquired as necessary. In our SILOD system, which is designed as a NLP system which has functions of MT, language identification, text compression etc. any AD that characterizes a specific language has a universal structure of dictionary items and special machine morphology. All the source language ADs have the same function and a united scheme organizations.

This scheme allows to unify such procedures of the source language text processing as a selection of minimum text units, the morphological analysis, the identification of the text with AD items, the organization of the dictionary information file.

Any lexical unit (LU) in AD acquires a description on the morphological, syntactic, semantic and functional levels as an appropriate characteristic set. The basic version of the system includes DI, which are defined as a set of the following characteristics:

- head LU as it is: a stem, a word-form or a MP;
- lexical and syntactic code (LSC), which depends on the typological features of the source language, its grammar and parsing or semantic analysis algorithms which are realized in the system in question;
- translation, which can be stored as a system of references to the corresponding target language items (stems and lexical and grammatical characteristics).

3. Linguistic Databases Software and Maintenance

When designing linguistic databases for practical NLP systems it is expedient to differ two types of bases:

- linguistic database (LDB), which include ADs in a convenient form and facilities of its updating and modification, which are oriented on the problems solved by the system designers (linguists, knowledge engineers etc.) and
- special dictionary files, which are results of special program-simulated conversion of the linguistic database into a format intended for the system software. As a result of such approach the format of the dictionary files can be changed as the system is developed, but the linguistic database (when the service utilities are advanced) can progress independently without obligatory rearrangements.

When designing a LDB for linguistic work we must take account of a linguists or a knowledge engineer tasks and to pay special attention to convenience of their work and easiness of LDB updating. Requirements to the production

ate and processing speed are not critical, the last must only orrespond to the rate of operator work on a computer. But ne convenience of the operator work is more than mportant.

When the DB is included in the NLP system the equirements to the production rate and processing speed ncrease critically. A major part of time such DB spends on ata look-up and reading. Addition of lexical information or orrection of dictionary items of LBD either is not performed r may be performed only if necessary as a special session.

Under these conditions it is not expedient to use as such latabase the LBD which has special functions of information ccumulation and modification and is to be convenient for a inguist and not the NLP system.

Теоретические принципы проектирования лингвистической базы данных для обработки естественного языка

Беляева Л.Н.

Резюме:
Обсуждаются квантитативные, лингвистические и компьютерные проблемы проектирования баз данных для систем обработки естестенного языка на примере баз данных для машинного перевода, реализованных в системе СИЛОД.

ANALYSING ORTHOGRAPHIC DEPTH OF DIFFERENT LANGUAGES USING DATA-ORIENTED ALGORITHMS

Antal van den Bosch[†] (Antal.vandenBosch@kub.nl), Alain Content[†] (acontent@ulb.ac.be), Walter Daelemans[‡] (Walter.Daelemans@kub.nl), and Beatrice de Gelder[*†] (degelder@kub.nl)

[†] Laboratoire de Psychologie Expérimentale, Université Libre de Bruxelles
[‡] Institute for Language Technology and AI, Tilburg University
[*] Department of Psychology, Tilburg University

SUMMARY

This paper proposes a quantitative operationalisation of the concept of orthographic depth, which plays a crucial role in the modelling of learning to read aloud in different languages. We propose to express the orthographic depth of a language by measuring (i) its complexity of letter-phoneme alignment, and (ii) its complexity of grapheme-phoneme correspondences. We presented (i) and (ii) as tasks to three data-oriented learning algorithms applied to English, French and Dutch data. Performance accuracy metrics are used to propose for each corpus a two-dimensional orthographic depth value.

Topical paper
Topic: automatic estimation of orthographic depth from text-to-speech corpora

INTRODUCTION

Within psycholinguistics, a growing interest is taken in comparing experimental data obtained with reading tasks across languages (e.g. see Katz & Frost, 1992). With respect to converting spelling to phonology, substantial differences exist between logographic, syllabic and alphabetic writing systems. Within the group of alphabetic writing systems, variations in language-specific influences on reading aloud have been characterized in terms of *orthographic depth* (Liberman, Liberman, Mattingly & Schankweiler, 1980; De Gelder, in press). The depth of an orthography can be understood as the degree to which it adheres to the alphabetic principle, i.e., the notion that spelling symbols have a one-to-one relation with phonemes. Orthographies in which there are more complex relations between letters and phonemes are described as deeper than more systematic spelling systems. More in detail, orthographic depth is often regarded as having two distinguishable aspects. The most important aspect relates to the complexity of the relations between the elements at the graphemic level (where *grapheme* relates to those letters or letter groups that map to single phonemes), and those at the phonemic level (phonemes). The second part relates to the diversity at the graphemic level, which is governed by language-specific graphemic, syllabic and morphological constraints (Klima, 1972; Liberman et al., 1980). Crosslinguistic experiments, based on lexical decision and naming latency measurements, suggest that the nature of psychological processes used to convert print into speech varies as a function of orthographic depth. More specifically, several authors have claimed that in shallow orthographies (e.g., Serbo-Croatian or Spanish), an analytic rule-based route (operating on grapheme-phoneme correspondences, GPCs) is used more than a lexical retrieval process (see e.g. Frost, Katz & Bentin, 1987).

The notion of orthographic depth has so far been dealt with informally (e.g., Coltheart, 1978; Katz & Frost, 1992); clearly, multi-lingual research could profit from a more precise operationalisation. Carello et al. (1992) tentatively claim that comparing rule-based GPC systems between languages may reveal differences in orthographic depth: Serbo-Croatian is likely to have a much smaller GPC set than, for example, English. An example of an actual construction of such a set is given by Coltheart et al. (1992), who present a model in which a GPC-set for English is learned from examples by a learning algorithm. As far as they do not require a priori language-specific constraints or heuristics, automatic, data-oriented learning seems to provide an appropriate means to extract statistical facts from language data relating to orthographic depth, without incorporating any linguistic bias. The application of data-oriented techniques for learning tasks like grapheme-to-phoneme conversion is language-independent, and can be applied to any language for which a corpus exists (Daelemans & Van den Bosch, 1993).

In this paper, we present three data-oriented learning algorithms which automatically learn to map graphemic strings to phonemic strings. We want to investigate whether application of these three algorithms to three alphabetic writing systems, viz. English, French and Dutch, reveal differences in orthographic depth between these three languages.

CORPUS SELECTION

We extracted our training and testing data from three computer-readable corpora of English, French and Dutch which consist of large lists of word spelling-pronunciation pairs. In the case of English, we used the *NETtalk* corpus of American English, first used by Sejnowski & Rosenberg (1987); the French data were extracted from the *Brulex* corpus (Content, Mousty & Radeau, 1990); the Dutch data were extracted from a large lexical data base made available for research purposes. A major concern of experimental validity was to obtain similarity between these corpora. This was done by restricting the size of the three corpora to about 20,000 words per corpus.

Since we are interested in, among other model features, the generalization capabilities of the three models, we split the three language data sets into training and test sets, in order for the models to be constructed or trained on the training sets, and tested for generalization performance on the test set. Each corpus was partitioned into a 1/13 test set (7.7% of the data set) and a 12/13 training set.

The training sets thus obtained consist of large lists of word-pronunciation pairs, for example, for English, the pair *shoe* - /ʃu/). For a system to be able to convert the 4-letter string *shoe* to the two-phoneme pronunciation /ʃu/, the system has to know that (i) the string *shoe* contains two graphemes, *sh* and *oe*, and that (ii) *sh* maps to /ʃ/, and *oe* maps to /u/ in this particular context. The knowledge needed for (i) is part of knowing which letter clusters can occur in a language; (ii) implies knowing what the possible correspondences between graphemes and phonemes within a language are. These two subproblems of converting spelling to pronunciation correspond to what is believed to be two seperate components of orthographic depth, i.e., (i) relates to complexity at the graphemic level, and (ii) relates to the complexity of the relation between the graphemic and the phonemic level. Furthermore, (ii) subsumes having solved (i).

Our experiments focused on analysing the complexity of (i) and (ii) separately. In the case of task (i), we presented a learning algorithm with the spelling-pronunciation pairs of the three training corpora. In the case of task (ii), we simulated the situation where (i) had already taken place, and trained two different learning algorithms on converting graphemic words to their phonemic transcription. In the case of English, these graphemic analyses were already available: in the NETtalk corpus, the phonemic strings are supplied with *phonemic nulls*, which are inserted at points where in the spelling string a graphemic letter cluster occurs. For example, the phonemic transcription of *shoe*, /ʃu/, is aligned as /ʃ-u-/. The same kind of alignment was performed for the Dutch and French corpora using pattern-matching algorithms and hand-correction. Surely, these algorithms and corrections introduce linguistic knowledge in a supposedly language-independent framework. A fully language- and linguistic knowledge- independent system would perform (i) and (ii) in sequence, using the graphemic analysis in (i) as input to system (ii). In fact, Daelemans & van den Bosch (submitted) demonstrate a data-oriented, language-independent system which integrates two high-performance data-oriented learning algorithms performing (i) and (ii) in sequence. In this paper, we focus on a seperate analysis of the two sub-problems.

THREE LEARNING ALGORITHMS

Grapheme-Phoneme Correspondences Extraction

An analysis of a spelling word into graphemes (i.e., letters or letter clusters mapping to a phoneme) primarily implies knowing which are the possible and typical graphemes in a language. The Grapheme-Phoneme Correspondences Extraction (GPCE) model described here is trained to capture this knowledge by an automatic, data-oriented learning algorithm. Rather than being trained explicitly on parsing a spelling string into graphemes, the GPCE model is aimed at constructing a memory base of hypothesised grapheme-phoneme mappings, so that after training the GPCE model is able to express probabilistic scores of a given graphemic analysis of a spelling word. The GPCE algorithm, in its most basic form, takes a training corpus of word-pronunciation pairs, and constructs on the basis of that corpus a memory base, containing all occurring grapheme-phoneme correspondences within that corpus. The GPC base construction algorithm has no knowledge of typical or regular grapheme-phoneme mappings: therefore, the graphemic analyses the algorithm comes up with may be linguistically impossible. To obtain this memory base of mappings, or rather *Grapheme-Phoneme Correspondence exemplars*, the following steps are taken for all word-pronunciation pairs in the training corpus:

(a) For each word-pronunciation pair, generate all possible parsings of the word in as much segments as there are phonemes (i.e., generate all possible letter clusters that can map to one phoneme). For example, the French word *chat* (cat), with pronunciation /ʃɑ/, results in three parsings: *cha t*, *ch at*, and *c hat*. (b) For each of the generated parsings, map each segment in that parsing to the corresponding phoneme. In the example of *chat*, this results in 6 GPC exemplars, two of which are correct (*): *cha*-/ʃ/, *ch*-/ʃ/ (*), *c*-/ʃ/, *t*-/ɑ/, *at*-/ɑ/ (*), and *hat*-/ɑ/. (c) For each derived GPC exemplar, store it in the GPC base. If it is already stored, increase the occurrence field of the GPC exemplar, and update the occurrence of the phonemic mapping (or create a new phonemic mapping field if the phonemic mapping was not encountered earlier). If it is not present in the GPC base, create a new exemplar, and initialise its occurrence field.

After training, a memory base is available which consists of a very large number of hypothesized GPC exemplars. The occurrence field of each of these GPC exemplars simply expresses the absolute number of occurrences of the GPC exemplar in the training corpus. The magnitude of this number is relative to two factors: (i) the size of the grapheme (single letter graphemes are encountered more frequently than multi-letter graphemes), and (ii) the probability of the grapheme. The algorithm described thus far has no direct relation with the problem of graphemic analysis, of which we want to investigate the complexity for the English, French and Dutch corpora. However, the memory base contains fuzzy information regarding the typicality of graphemes. This knowledge can be used to estimate for new, unseen test words their most probable graphemic analyses. To obtain these estimates, the following algorithm is implemented: for each unseen test word, (a) generate all possible graphemic analyses. On the one end, an analysis is generated which takes each letter as a seperate grapheme; the other extreme is an analysis containing only graphemes of maximal length (e.g., 4 letters in English, as in *ough*); (b) for each graphemic analysis, search the GPC exemplar base for all matching GPC exemplars. Each analysis is given a score which is the sum of the occurrences of all matching GPC exemplars; (c) the analysis which is assigned the highest score, is taken as output. Given the analysis already present in the prepared corpus, it can be determined for each test word if the graphemic analysis is correct. This model feature is examined in the Results section.

Trie Compression

A detailed description of the Trie Compression algorithm can be found in Van den Bosch & Daelemans (1993). The Trie Compression algorithm automatically stores grapheme-phoneme knowledge into a tree-like memory structure (the Trie) in such a way that grapheme-phoneme correspondences are stored with the exact amount of contextual knowledge that makes the mappings unambiguous. A correspondence is stored in the Trie in the form of a *path* through the Trie, leading past all occurring context letters, and ending in a Trie node denoting the unambiguous phonemic mapping. 'Irregular' mappings, i.e., graphemes that need large contexts in order to disambiguate between possible phonemic mappings, are stored further ('deeper down') in the Trie, where 'depth' of storage of a grapheme-phoneme mapping is directly related to the amount of context graphemes needed to disambiguate between similar graphemic string mapping to different phonemes. The knowledge present in a training corpus is stored in the Trie in such a way that no grapheme-phoneme correspondence information is lost. At the same time, the amount of memory needed to store this information is minimized. This lossless compression ensures that for any phonemic mapping in the training corpus, there is a path in the Trie leading to an end node containing unambiguous phonemic information. However, this may not be the case for a test word which might contain substrings not encountered in the training corpus. When Trie Search is performed with such a string, the retrieval algorithm will not be able to retrieve unambiguous phonemic information. The system attempts to solve this problem by storing at every node information about the *most probable* phonemic mapping. When search fails, this extra information enables the model to always come up with a 'best guess', a property of the model which is essential for optimizing generalisation performance.

For each of the three language corpora, the amount of compression compared to the original training data as well as the performance accuracy scores on test material will be examined more closely in the Results section.

Similarity-Based Reasoning

The Similarity-Based Reasoning (SBR) model attempts, just as the Trie Compression model, to store grapheme-to-phoneme mapping knowledge in such a way that it can be successfully used to retrieve the phonemic transcription of new, previously unencountered test words.

During training of the SBR model, a *memory base* is constructed consisting of *exemplars*, which in the case of grapheme-to-phoneme mappings consist of patterns of strings of graphemes (one focus grapheme surrounded by context graphemes), and the associated phonemes and their distribution in the training corpus (as there may be more phonemic mappings to a graphemic string). To construct the memory base, each word in the training corpus is converted in a number of patterns. Each pattern consists of a focus grapheme surrounded by a fixed number of left and right context graphemes, together with the associated phoneme. For our models, we kept the number of left and right context characters at 5. Patterns are stored as exemplars in the memory base; whenever a duplicate graphemic pattern is found, the occurrence count of the phonemic mapping of the stored exemplar is increased, or a new phonemic mapping field is added to the exemplar if that phonemic mapping was not encountered earlier. To retrieve the phonemic transcription of a test word, it is converted into patterns. Each of these patterns is matched against all memory exemplars. If the test pattern matches an exemplar, the category with the highest frequency associated with it the exemplar is retrieved. If it is not in memory, all memory items are sorted according to the similarity of their pattern to the test pattern. The (most frequent) phonemic mapping of the highest ranking exemplar is then predicted as the category of the test pattern.

A more detailed description of the SBR algorithm can be found in Daelemans & Van den Bosch (1992).

RESULTS

Grapheme-Phoneme Correspondences Extraction

The GPC memory base construction algorithm was applied to training sets of French, English and Dutch which are a subset of the original training set. Each training set contained 5000 words. This smaller set was chosen, since pilot experiments indicated a convergence of performance at data set sizes above approximately 1000 words. After construction, the full original training set was processed through the GPC test algorithm. From the resulting best guessed graphemic analyses and phonemic mappings, performance scores were computed expressing the percentage of incorrect graphemic analyses of words. Table 1 lists these figures for the three languages.

corpus	% incorrectly aligned words
English	75.5
French	87.1
Dutch	78.7

Table 1. *Number of incorrectly aligned test words obtained with the three GPCE models after memory base construction, trained on 5000-word partitions of the original training sets and tested on the full test sets.*

It is obvious that the performance scores listed in Table 1 are not high. This is due to the overall fuzziness of the GPCE model, being mainly concerned with storing regularity by only being sensitive to frequency. More importantly, the differences between the three corpora are apparent. In the case of the English corpus, alignment is relatively less complex than in the cases of the Dutch and the French corpus. In terms of correctly aligned words, the French model clearly renders the least accurate results.

Trie Search

The application of Trie Compression to the three training corpora resulted in three models of very different size. Since Trie Compression is based on removing redundancy from a corpus, higher compression indicates that the corpus contains more redundancy (i.e., more regularities). In terms of compression of memory usage, the French model was compressed by a factor of 90.8%, the Dutch model by 87.4%, and the English model by 70.9%. The English data appears to contain less redundancy, and can be regarded as more irregular than the French and Dutch data. The performance on the test data provides more clues concerning differences between English on the one hand and French and Dutch on the other. Table 2 lists the generalisation performance of the three models on the test data.

language	% incorrect words	% incorrect mappings
English	45.7	9.0
French	10.9	1.7
Dutch	18.6	2.4

Table 2. *Generalisation performance (on test data) of the three models. Scores listed on incorrectly produced words and incorrectly transliterated letters.*

Best performance scores are obtained with the French model. In terms of correctly transliterated words, the Dutch model scores significantly lower, but in terms of correctly converted phonemes (the most unbiased measure), the scores are

roughly similar. The English model scores noteably worse than the French and Dutch model on both words and phonemes.

Similarity-Based Reasoning

As described earlier, the SBR memory base was constructed for each corpus by converting all word-pronunciation pairs into fixed-length window patterns, which were then stored as exemplars in the memory base. Since there were not many duplicate 5-1-5 patterns in any of the three corpora, large memory bases resulted. For example, in the case of English, out of the 135,406 5-1-5 patterns 120,062 exemplars were stored (11.3% compression). For Dutch, compression was 12.0% (156,449 exemplars stored), and for French 17.8% (129,054 exemplars stored), indicating the fact that the French corpus contains more partly similar words than the other two corpora.

Table 3 displays the generalisation accuracy on test words and phonemes for the three models. The results show high scores for Dutch and French, and a significantly lower score for English, especially when expressed in the percentage of correctly transliterated words. The performance results are highly similar to those obtained with the Trie Search models.

language	% incorrect words	% incorrect mappings
English	45.9	9.0
French	11.0	1.7
Dutch	17.5	2.2

Table 3. *Generalisation accuracy on test words and mappings by the three Similarity-Based Reasoning models.*

CONCLUSIONS

The application of three data-oriented machine learning techniques on three grapheme-to-phoneme corpora has revealed differences between the orthographic complexity within these corpora. In line with the propositions of Klima (1972) and Liberman et al. (1980) we propose that the problems of graphemic alignment and grapheme-to-phoneme conversion are the basic components of converting spelling words to their phonemic transcription. Although they are not totally independent problems, they can be regarded as the two most distinct components or dimensions in the space describing the complexity of an orthography.

We argued earlier that the first dimension of orthographic depth, i.e., the complexity of graphemic analysis (i.e., the problem of aligning letter strings to phonemic strings), is embedded in the memory base of the GPCE model, and is expressed in the error output of the model when applied to unseen test words. Tables 1, 2 and 3 display the differences obtained between the GPCE models of the three language corpora. We propose to take the measure indicating the number of incorrectly aligned words to express the complexity of dimension (i) of orthographic depth. It should be stressed that the absolute magnitude of the measures is not important here: the key importance lies in the *relative differences* between the three languages.

The complexity of converting strings of graphemes to strings of phonemes is, amongst other measures like Trie Compression factors, Trie sizes and SBR memory base compression factors, most prominently expressed in the generalisation accuracy on the production of phonemes in test words. Furthermore, Trie Compression and SBR performances are highly similar (see Tables 2 and 3). We propose to take this performance measure as a measure of complexity of going from the level of graphemes to the level of phonemes, i.e., dimension (ii) of orthographic depth: the higher this number is, the more complex the problem is within a certain corpus. Again, only the relative differences between the three languages matter here.

The two dimensions and the three points marking the three corpora are displayed graphically in Figure 1, constituting a 'map' in which the relative distance of the three corpora within the two-dimensional orthographic depth space is clearly expressed.

Our data-oriented, generic, two-dimensional classification of the complexity of grapheme-to-phoneme conversion can be used as a platform for determining an unbiased grounding of orthographic depth for any corpus in any language. The only restriction the corpus must adhere to is the approximate number of words. In our opinion, our presupposition of á number of approximately 20,000 words is sufficiently large.

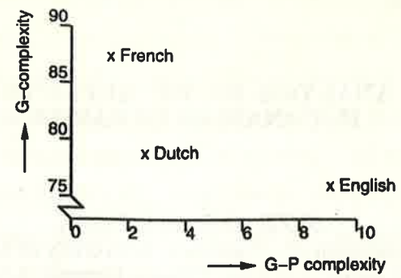


Figure 1. *Graphical display of the two-dimensional orthographic depth space, with 'x's marking the three corpora.*

Acknowledgements

This research was partly supported by a grant from the Human Frontier of Science Programme *Processing consequences of contrasting language phonologies.* We would like to thank Terrence Sejnowski and Henk Kempff for making available for research purposes the NETtalk corpus and the Dutch corpus, respectively.

REFERENCES

Bosch, A. van den & W. Daelemans (1993). Data-oriented methods for grapheme-to-phoneme conversion. *Proceedings of the European Chapter of ACL*, Utrecht, 45-53.

Carello, C., M.T. Turvey, & G. Lukatela (1992). Can theories of word recognition remain stubbornly nonphonological? In *Haskins Laboratories Status Report on Speech Research 1992*, 193-204.

Coltheart, M. (1978). Lexical access in simple reading tasks. In G. Underwood (Ed.), *Strategies of information processing*, 112-174. New York: Academic Press.

Coltheart, M., B. Curtis, P. Atkins & M. Halter (in press). Models of reading aloud: Dual-route and Parallel-Distributed-Processing approaches. *Psychological Review*, in press.

Content, A., P. Mousty, & M. Radeau, (1990). Brulex: une base de données lexicales informatisée pour le francais écrit et parlé. In L'Année Psychologique, **90**, 551-566.

Daelemans, W. & A. van den Bosch (1993). TabTalk: Reusability in data-oriented grapheme-to-phoneme conversion. In *Proceedings of Eurospeech '93*, Berlin.

Daelemans, W. & A. van den Bosch (submitted). A language-independent, data-oriented architecture for grapheme-to-phoneme conversion.

De Gelder, B. (in press). Reading acquisition: The rough road and the silken route. *Journal of Chinese Linguistics.*

Frost, R., L. Katz & S. Bentin (1987). Strategies for visual word recognition and orthographical depth: a multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, **13**:1, 104-115.

Glushko, R.J. (1979). The organisation and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, **2**, 361-379.

Katz, L. & R. Frost (1992). The reading process is different for different orthographies: the orthographic depth hypothesis. In *Haskins Laboratories Status Report on Speech Research 1992*, 147-160.

Klima, E.S. (1972). How alphabets might reflect language. In J.F. Kavanagh & I.G. Mattingly (Eds.), *Language by ear and by eye: The relationship between speech and reading.* Cambridge, MA: MIT Press, 57-80.

Liberman, I.Y., A.M. Liberman, I.G. Mattingly, & D.L. Shankweiler (1980). Orthography and the beginning reader. In J.F. Kavanagh & R.L. Venezky (Eds.), *Orthography, reading and dyslexia*, Baltimore: University Park Press, 137-153.

Seidenberg, M.S. & J.L. McClelland (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, **96**:4, 523-568.

Sejnowski, T.J. & C.R. Rosenberg (1987). Parallel networks that learn to pronounce English text. *Complex Systems*, **1**, 145-168.

VARIABLE RULE ANALYSIS OF V/U ALTERNATION CONSTRAINTS IN CANADIAN UKRAINIAN

SVITLANA BUDZHAK
Department of Linguistics, University of Ottawa
78 Laurier Ave East, Ottawa, Ontario, K1N 6N5
Canada; e-mail: 058372@acadvm1.uottawa.ca

Summary: Multivariate analysis has been applied to examine the conditioning of linguistic and extralinguistic environments on v/u alternation in one of the Eastern Slavic languages, Ukrainian. A computerized program makes it possible to account for various effects simultaneously.The obtained result indicates that this variation is mostly constrained by the underlying position of the variable in a syllable structure as well as the phonological segment following it.

Topical paper in sociolinguistics

V/u alternation in Ukrainian exemplified in (1) has traditionally been explained as a result of the tendency dominating in a certain dialect, idiosyncratic pronunciation and the nature of the adjacent sounds (Zilyns'kyj 1979, Cerkevich & Pavlovsky 1982). Ukrainian orthography dictates V-variation only for prepositions and prefixes (Ditel' 1990, Humensky 1980), though it was also noted that [v] could be articulated as [u] within or at the end of the word (Zilyns'kyj 1979). The purpose of this study is to investigate the variable patterns of this alternation and to establish the constraints conditioning it. We will also offer an explanation for the nature of this phenomena, which for some reason surfaces only in this Eastern Slavic language.

(1) a. **Naš** **učytel'** vs. **naša** **včytel'ka**
our(masc) teacher(masc) our(fem) teacher(fem)

b. **Učora** vs. **včora**
yesterday

c. **U** nas vs. **v** nas
at ours

The theoretical background for this research has been adopted from the framework of Variation theory (Labov 1966, 1972, Poplack 1980, 1988, Sankoff 1988). This involves the scientific investigation of language use manifested in its natural context, and the statistical quantitative analysis of variation in linguistic forms illuminated by features of their linguistic and extra-linguistic environments (Sankoff 1988).

Variation in the use of linguistic variants can be influenced by certain features of their phonological, syntactic, semantic or discourse contexts as well as sociodemographic characteristics of speakers. Therefore, we assume that both sociolinguistic and linguistic

factors condition **v/u** alternation in Ukrainian. Since this alternation is mandatory by prescriptive grammar rules only for prepositions and prefixes (Ditel' 1990), we shall ascertain whether educated people show different rates and/or patterns of variability than those with less formal education. Speech style may also play a role. Among purely linguistic factors we consider phonological, syntactic and lexical environments. Our hypothesis is that either [**u**] is a morphophonemic variant of [**v**] or [**v**] is a morphophonemic variant of [**u**] irrelevant of its morphological status (i.e. preposition, prefix, etc.), since their morphosyntactic roles are identical and depend only on the phonological environment according to the prescriptive grammar.

The data for this research was extracted from sociolinguistic interviews with four Ukrainian Canadians, which were held in Ukrainian and are approximately 8 hours long. All informants were randomly selected using the social network techniques (Labov.1966). Two elderly speakers represent the first generation, i.e. they were brought up in Western Ukraine and left the country in their early twenties. The two younger informants were born in Canada and represent the second generation. Both men have a post-secondary education (i.e. holders of University degrees), whereas only one woman has been studying at the University.

All tokens containing the circumscribed variable were extracted from the corpus. Tokens of [**v**] in the onset immediately followed by a vowel were excluded from the data, since no variation occurred in this context (37/37). There was almost no variation (117/120) observed in the data of the second generation, which did not allow us to include it into our analysis. Therefore, the remaining corpus consists of 332 tokens of the variable under study.

All tokens were analyzed by GOLDVARB 2.0, a variable rule (VR) application for the Macintosh (Rand & Sankoff 1990). Variable rule analysis is a statistical procedure which extracts regularities and tendencies from data presumed to have a random component. Maximum likelihood estimation based on logistic regression carries out the quantitative computation of choices "repeated many times in a variety of contexts, each context being defined as a specific configuration of conditioning factors (Sankoff 1988).

Each lexical item containing a potential context for **v/u** alternation was coded according to each of seven extra-linguistic and ten linguistic factor groups. The extra-linguistic factors include speaker's age, sex, socio-economic class, education, native language, linguistic market membership (Sankoff and Laberge 1978) and speech style. Purely linguistic factors include: preceding and following phonological segments, preceding and following stress, underlying position in the syllable, morphological status of the variable, preceding and following grammatical category, following Case and the part of speech containing the variable.

Table 1. **Contribution of factors selected as significant by the VR program to the probability that [v] will be realized as [u]**

Corrected mean: .401

Following phonological segment		Underlying position in the syllable	
Consonant [V]	0.735		
Affricate	0.703	Rhyme	0.563
Sonorant	0.584		
Stop	0.553	Onset	0.406
Glide	0.404		
Sibilant	0.357		
Vowel	0.289		
Pause	0.280		

Factors not selected: preceding phonological segment, preceding and following stress, morphological status of the variable, preceding and following grammatical category, following Case, part of speech.

The stepwise multiple regression procedure incorporated in the VR program retains only those factors that contribute a statistically significant effect to the probability that the variable "rule" in question will be applied; in this case that **[v]** will be realized as **[u]**. Table 1 shows only two linguistic factor groups to be significant for the probability of **[u]** realization: 1) the following phonological segment and 2) underlying position in the syllable. None of the extra-linguistic factors significantly conditions the variation.

The factor group of following phonological segment was designed to roughly reflect the sonority hierarchy, which ranks the sounds according to the degree of openness of the vocal apparatus (see Goldsmith 1989 for discussion). It is well established that the sonority hierarchy plays an important role in syllabification, i.e. the more sonorous the sound, the nearer to the peak of the syllable it appears, and vice versa, the less sonorous - the further. Therefore, it is interesting to notice that the sonority hierarchy is largely observed in our results. This means that when the variable is followed by a consonant the probability of **[u]** occurring is higher than when it precedes a vowel. In the first case syllabification will take place, whereas in the second one it will not. The finding that the only other significant factor group is underlying position of the variable in the syllable supports our assumption. Hence, the obtained results leads us to conclude that **v/u** alternation is conditioned by the process of syllabification.

Using the same computational program it was possible to cross tabulate the interaction of different factor groups in order to justify the obtained results. In our case it is necessary to explain the possible violation of sonority hierarchy by sibilants which occur between glides and vowels, and stops which appear slightly higher than sonorants (.553

and .584 correspondingly). It is evident from Table 2 that in 86% of the (70) cases where the variable was followed by a sibilant, it was also preceded by a vowel. This suggests that the variable was syllabified as a coda since that position was available, whereas the onset was originally occupied by a sibilant, and hence there was no need for alternation.

The distribution of the data also explains the sonority violation by stops. In 83% of (95) cases where the variable occurred in front of a stop, it was also preceded by a vowel. This implies that the variable has a higher probability of being, explaining why the [v] variant is favored (63%) here.

Table 2. **The interaction of the preceding and following phonological environments in v-variation.**

Following segment		Preceding segment: Vowel N	%	Sonor. N	%	Stop N	%	[V] N	%	Pause N	%	Sibil. N	%	**Total** N	%
Sibilant	**V**	45	75	3	100	2	67	0	-	1	100	3	100	54	77
	U	15	25	0	0	1	33	0	-	0	0	0	0	16	23
		60		3		3		0		1		3		70	
		(86%)		(4%)		(4%)		0		(1%)		(4%)		(100%)	
Stop	**V**	50	63	0	0	0	0	4	57	0	0	3	75	57	60
	U	29	37	2	100	2	100	3	43	1	100	1	25	38	40
		79		2		2		7		1		4		95	
		(83%)		(2%)		(2%)		(7%)		(1%)		(4%)		(100%)	

The other significant factor constraining **[u]** realization is the underlying position of the variable in the syllable. The probability of **[u]** occurring in the rhyme is higher (.563) than in the onset (.406) (see Table 1), while the opposite is the case for **[v]**. Note that according to the prescriptive grammar **v/u** alternation takes place only at the beginning of the word or in prepositions, whereas the obtained results show that in reality the probability for the alternation to take place is higher when [v] is at the end of the syllable or word. This can be explained by the theory of syllable weight (Zec 1988).

Since Ukrainian is in the transitional area between Polish and Russian, it shares with the West the potential for a two-position syllable rhyme, whereas like Eastern Slavic dialects, it is restricted to only one mora. Therefore, [V] in the rhyme may be associated with the redundant moraic tier and become [-cons] since diachronically it reflexes Common Slavic mora-carrying liquid (Bethin 1993), but it does not contribute to the syllable weight and therefore, on the surface it is realized [+cons] (see Budzhak 1993 for more detail).

Thus, we can conclude that **v/u** alternation in Ukrainian is mainly constrained by the phonological environment and the structure of the syllable. The nature of the following

phonological segment as well as the relation of the variable to syllabification processes influence this variation considerably. If [v] is followed by the consonant [V] itself or an affricate the probability for its vocal counterpart to occur is the highest, whereas before vowels and pauses it is the lowest. This result mitigates against some established orthographic rules on **v/u** alternation (see Ditel' 1990).

The results obtained in the research help to clarify the traditional explanations of **v/u** alternation, and the origins of this phenomenon. It proves that neither of extra-linguistic factors (f.g. education, age, etc.) conditions the variation. The realization of the variable primarily depends on the syllabification process, i.e. to which element on the CV-tier the variant will be associated with (Clements & Keyser 1983). If [v] is independent of word structure (i.e. it is a preposition), it is syllabified across the word boundary in the lexical form, whereas if it occurs within a word, it may become [-consonantal] in the process of resyllabification or being associated with a moraic tier in the case of a rhyme. Hence, the results of the research support our assumption that there is only one preposition in Ukrainian with two morphophonemic variants [**u**] and [**v**].

REFERENCES:

Budzhak, S. 1993. "*V-variation in Ukrainian and its syllabification*". Ms. University of Ottawa.

Clements, G. and Keyser, S. 1983. "*CV-Phonology*". LI Monograph Series, no.9. Cambridge, Nass.: MIT Press.

Cerkevich, K. and Pavlovsky, V. 1982. "*Ukrainian Language Reference book*", volume 1. Research Society for Ukrainian Terminology, New York.

Ditel', O.(ed.), 1990. "*Ukrajins'kyj pravopys*". Kyjiv: Naukova dumka.

Goldsmith, J. 1989. "*Autosegmental and Metrical Phonology*". Oxford, UK: Basil Blackwell.

Humesky, A. 1980. "*Modern Ukrainian*". Canadian Institute of Ukrainian Studies, Edmonton.

Labov, W. 1966. "*The Social Stratification of English in New York City*". Washington, D.C.: Center for Applied Linguistics.

Labov, W. 1972. "*Sociolinguistic Patterns*". Philadelphia: University of Pennsylvania Press.

Poplack, S. 1980. "Sometimes I'll start a sentence in Spanish y termino en espanol". *Linguistics* 18:581-618.

Poplack, S. 1988. "Language status and language accommodation along a linguistic border". In P.Lowenberg (ed.), *Georgetown Universitu Round Table on Languages and Linguistics 1987*. Washington, DC: Georgetown University Press. 90-118.

Sankoff, D. 1988. "Variable rules". In U.Ammon et al. (eds.) *Sociolinguistics: an International handbook of the science of language and society*. Berlin: Walter de Gruyter, 784-997.

Sankoff, D. and Laberge, S. 1978. "The linguistic market and the statistical explanation of variability". In Sankoff, D. (ed.) *Linguistic variation: models and methods*. New York , 239-250.

Zilyns'kyj, I. 1979. "*A Phonetic Description of the Ukrainian Language*". Harvard University Press.

Corpus-Based Analyses of Adjectives: Automatic Clustering

Kuang-hua Chen and Hsin-Hsi Chen*

Department of Computer Science and Information Engineering

National Taiwan University

Taipei, Taiwan, R.O.C.

e-mail: hh_chen@csie.ntu.edu.tw

Abstract

Similarity analysis is a substantial issue in both corpus-based researches and language usages. This paper focuses on the semantic usages of adjectives, and analyzes the similarities among adjectives. The adjective and the semantic tag of the head noun that it modifies in a noun phrase form a co-occurrence. A two-stage algorithm is applied to clustering the adjectives according to these co-occurrence relationships. Experimental results show that we break even the two issues of large data clustering and meaningful clustering.

Paper Category: Topical Paper.

Topic Area: Corpus Linguistics, Similarity Analysis, Clustering.

1. Introduction

Since the importance of real-world applications is committed in recent years, corpus-based researches become the core of the field of computational linguistics. Many models, such as hidden Markov model, word association model, cache-based model, *etc.*, have been proposed to deal with practical applications. An important problem in these models is how to calculate the reliable probabilities of events. Many smoothing methods are reported. Most of these methods use low degree Markov probabilities to replace the unreliable high degree probabilities. A thorough resolution to this problem is to cluster the events. Brown *et al.* [1] propose class-based language models to group words directly on the training table. Clustering not only reduces the memory needed in corpus-based tasks, but also smoothes the probabilities of events. In addition, word groups could be further investigated for language usage and lexicography. A method to group nouns according to the predicate-argument structures is described by Hindle [2]. Hatzivassiloglou and McKeown [3] analyze the co-occurrences of adjectives and nouns, and then cluster the adjectives. All these methods investigate the relationships among word surface forms. In this paper, we intensify the semantic usages of words. The idea is to examine the noun phrases in text corpora, and assign

* To whom all the correspondences should be sent.

semantic tags to head nouns. Then, the co-occurrences of the premodifying adjectives and the semantic tag of head noun provide the clues for clustering.

2. The Work

The proposed method uses a probabilistic chunker [4] to generate chunked texts, and determines the possible boundaries of phrase structures. All the noun chunks that contain adjectives are considered for further analyses. The head nouns of these noun chunks are assigned semantic tags, and then the relationships of adjectives and the semantic tags are investigated. The semantic tags of nouns are defined by Roget's Thesaurus. Roget's Thesaurus defines 1000 tags, and these tags are the leaves of a tree-like structure. Six classes are given. Various sections are defined under the classes and the 1000 tags are characterized. Table 1 shows the plan of classification of Roget's Thesaurus.

Table 1. Classification of Roget's Thesaurus

CLASS	SECTION	TAG	CLASS	SECTION	TAG
ABSTRACT RELATIONS	Existence	1 - 8	SPACE	In General	180 - 191
	Relation	9 - 24		Dimensions	192 - 239
	Quantity	25 - 57		Form	240 - 263
	Order	58 - 83		Motion	264 - 315
	Number	84 - 105	MATTER	In General	316 - 320
	Time	106 - 139		Inorganic	321-356
	Change	140 - 152		Organic	357 - 449
	Causation	153 - 179	AFFECTIONS	In General	820 - 826
INTELLECT	Formation of Ideas	450 - 515		Personal	827 - 887
	Communication of Ideas	516 - 599		Sympathetic	888 - 921
VOLITION	Individual	600 - 736		Moral	922 - 975
	Intersocial	737 - 819		Religious	975 - 1000

The overall analysis procedures are summarized in Figure 1.

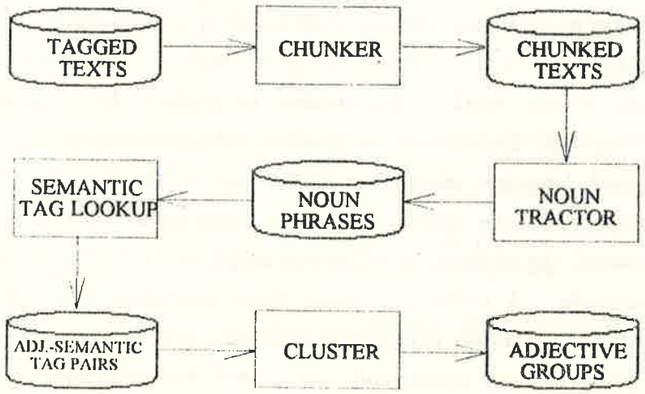


Figure 1. Experimental Procedures

The test texts are LOB Corpus which contains 9123 different adjectives. The component CHUNKER is trained from LOB Corpus underlying bigram language model, i.e., two probabilities are considered. One is the probability of a chunk; the other is the probability of a chunk given its previous chunk. The NOUN-TRACTOR, which is a finite state mechanism containing 9 states, extracts the maximum length noun phrases from chunks sequence. The chunker has 98% chunk correct rate and 94% sentence correct rate in outside test [4]. The average precision of NOUN-TRACTOR is 95% [5]. Due to the high performance of these two components, the extracted noun phrases are suitable for clustering the adjectives.

Two similarity modules are considered in Hatzivassiloglou and Mckeown [3], one is the co-occurrences of adjectives and nouns; the other is those of adjectives and adjectives. Their postulation is: it is impossible that two adjectives modifying the same nouns belong to the same cluster. However, it is not true in many real examples. Some of these shown as follows are quoted from LOB Corpus.

(1) The two *rival African nationalist* parties of Northern Rhoodesia. (A01:25)
... crochet and tatting in *fine* and *medium*-weight cottons ... (E01:55)
... an electrical drill, *pure* and *simple* ... (E03:104)
They are very *simple*, *cheap* and *easy* to make. (E04:97)

As the result, we do not use the negative evidences for similarity analysis. Currently, only co-occurrences of adjectives and noun tags are considered.

3. Clustering

Our vocabulary consists of 9123 different adjectives and 1001 possible semantic tags for nouns (an extra semantic tag is used for unknown words). To cluster the 9123 adjectives is intractable in many clustering methods. The method proposed by Hatzivassiloglou and McKeown [3] costs much computing time. This is why only 21 different adjectives are used as test set in their paper. Here, a two-stage clustering algorithm [6] is employed to cluster the adjectives. The co-occurrences of adjectives and semantic tags can be regarded as a matrix (say original matrix, OM) with 9123 rows and 1001 columns. Each entry indexed by (i,j) in the matrix is the frequency of co-occurrence of the i'th adjective and the j'th semantic tag in the testing corpus.

According to the entries in OM, a bit matrix BM is generated under the following rule.

(2) For each entry (i,j) in OM, if $OM(i,j) > 0$, then set $BM(i,j)$ to 1.
Otherwise, set $BM(i,j)$ to 0.

The similarity of two adjectives is measured by the respective row vectors in BM. (3) gives the similarity measure. $RI[k]$ denotes the k'th element of row vector and $\oplus$ denotes the exclusive or.

(3) $Sim(RV_{r_i}, RV_{r_j}) = 1001 - \sum_{k=1}^{1001} RV_{r_i}[k] \oplus RV_{r_j}[k]$

The high similarity measure means the two row vectors are highly similar to each other. Based on the similarity measure, an optimal row index sequence (*ORIS*) is generated. The first element of the *ORIS* is obtained by choosing the row vector indexed by r_i which has the highest similarity measure with zero vector. Then, the second element is obtained by choosing the row vector index r_j which has the highest similarity measure with RV_{r_i}. The rest elements can be obtained in the same way. The above steps form the first stage of the clustering algorithm. ORIS will guide the clustering procedure in the next stage, and will reduce the complexity of the overall algorithm.

The second stage finds the clusters according to a predefined threshold value and the *ORIS* generated in the first stage. The threshold value determines how many clusters will be generated. At first, the zero row vectors in *OM* are clustered into an initial class RC_0, say m_0 zero vector. Assume *k-1* clusters, RC_1, RC_2, ..., and RC_{k-1}, are formed, and these clusters consist of m_1, m_2, ..., and m_{k-1} adjectives, respectively. Let $m = m_0 + m_1 + m_2 + \ldots + m_{k-1}$. Initiate a new cluster RC_k with only one row vector $RV_{ORIS_{m+1}}$. Add $RV_{ORIS_{m+2}}$, $RV_{ORIS_{m+3}}$, ..., into the new cluster RC_k until the information loss is larger than the predefined threshold. The information loss is defined below.

(4) $IL_k = \sum_{i=1}^{m_k} \sum_{j=1}^{1001} abs(RV_{ORIS_{m+i}}[j] - RV_{RC_k}[j])$

where $RV_{RC_k} = \sum_{i=1}^{m_k} RV_{ORIS_{m+i}}$, and function *abs* returns an absolute value.

The complexity is shown to be $O(M^2N)$ for an $M \times N$ matrix, so that the complexity of this clustering algorithm is tractable for large vocabulary application like our work.

4. Experimental Results

The first experiment we conduct is a repetition in [1], i.e., cluster the 21 adjectives listed in Table 2. Because the adjective *antitrust* does not occur in the test LOB Corpus, we exclude it in our experiment.

Table 2. Adjectives Used in Experiment I

antitrust	big	economic
financial	foreign	global
international	legal	little
major	mechanical	new
old	political	potential
real	serious	severe
staggering	technical	unexpected

Table 3 demonstrates the experimental result. In general, the clusters correspond to the common usages. The adjectives *old* and *new* are grouped into the same cluster, not two different clusters shown in [1]. This is because the negative evidences are not used in our experiment. From the viewpoint of language usage, *new* and *old* are used to modify the same kind of nouns. As the result, they belong to the same group in our model.

Table 3. Experimental Results of Experiment I

Cluster	Words
1	economic, financial, unexpected, potential, legal
2	mechanical, technical, international, foreign, global,
3	major, serious, severe
4	real, big
5	political, old, new
6	little, staggering

Economic and *financial* are grouped together, but *political* are included into other cluster. These words are expected to be in the same cluster. However, due to the limited size of LOB Corpus (only 1M words), some bad clusters is unavoidable.

The second experiment we consider is to cluster some hyphened adjectives selected from test corpus. Total 21 adjectives are included in the experiment and segmented into 12 clusters. These words and the clusters to which they belong are listed in Table 4.

Table 4. Some Results of Experiment II

Cluster	Words
1	public-opinion
2	American-Indian, British-Caribbean
3	best-known, little-known, present-day
4	important-looking, politico-economic
5	main-line, right-angled
6	white-armed, younger
7	whole-hearted
8	long-ago
9	small-bowled, large-scale
10	old-age, new-born
11	good-humoured
12	great-power, high-backed

The clusters in Table 4 show many uncommon hyphened adjectives are also grouped into some meaningful clusters.

The last experiment is to cluster all of the adjectives in the test corpus, i.e., the 9123 adjectives. The resulting clusters are 367. This experiment takes about 24 hours. Because many low-frequency adjectives involve in, the experimental result is heavily disturbed. The meaningful clustering is not attained in the global viewpoint. But we still receive some good clusters. For example, the cluster 256 consists of *Nazi-style*, *socialistic*, *nationalistic* and

prohibitive. Cluster 342 consists of *German-French*, *French-Canadian*, *American-Indian*, *British-Caribbean* and *Soviet-American*.

5. Concluding Remarks

For reliable estimation of probabilities in corpus-based researches, clustering is indispensable. From viewpoint of language usages, to cluster words is a good way for comparing words. Usually, the researches in grouping words focus on the surface forms of words, and try to find the implicit lexical-semantic relationship. In this paper, we cluster adjectives according to their semantic usages directly. The associations of the semantic tags of head nouns and their adjective modifiers are considered. Namely, a link is built between word and semantic tag, not just word and word. Since clustering is an NP problem, experiments on large volume of data seem to be intractable. However, practical applications are important and unavoidably this kind of researches involve very large data. To make the clustering for large data tractable, a two-stage clustering algorithm is applied. Comparing to 21 adjectives tested in Hatzivassiloglou and McKeown [3], 9123 adjectives occurring in the LOB Corpus form the test set. Due to the training size of LOB Corpus, the results shown in our work demonstrate both good clusters and bad clusters. Very large corpus should be used to prove the effectiveness. Another problem is how to assign a unique semantic tag to head noun. Future works should focus on these two issues.

Acknowledgments

We are thankful to Ren-Feng Chang and Yue-Shi Lee for their helps in programming.

References

[1] P.F. Brown, V.J. Pietra, *et al.*, "Class-Based N-Gram Models of Natural Language," *Computational Linguistics*, 18(4), 1992, pp. 467-479.

[2] D. Hindle, "Noun Classification from Predicate-Argument Structures," *Proceedings of 28th Annual Meeting of ACL*, 1990, pp. 268-275.

[3] Vasileios Hatzivassiloglou and Kathleen McKeown, "Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning," *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 1993, pp. 172-182.

[4] K.-H. Chen and H.-H. Chen, "A Probabilistic Chunker," *Proceedings of ROCLING VI*, 1993, pp. 99-117.

[5] K.-H. Chen and H.-H. Chen, "Extracting Noun Phrases from Large-Scale Texts: A Hybrid Approach and Its Automatic Evaluation," Submitted to *ACL94*.

[6] H.-H. Chen and Y.-S. Lee, "An Unsupervised Clustering Algorithm for Storage Reduction in Corpus-Based Applications," Submitted to *COLING94*.

APPROXIMATE N-GRAM MARKOV MODEL FOR NATURAL LANGUAGE GENERATION

Hsin-Hsi Chen and Yue-Shi Lee

Department of Computer Science and Information Engineering

National Taiwan University

Taipei, Taiwan, R.O.C.

E-mail: hh_chen@csie.ntu.edu.tw

Abstract

This paper proposes an Approximate n-gram Markov Model for bag generation. Directed word association pairs with distances are used to approximate (n-1)-gram and n-gram training tables. This model has parameters of word association model, and merits of both word association model and Markov Model. The training knowledge for bag generation can be also applied to lexical selection in machine translation design.

paper category: topical paper; **topic area:** application of models, finding of NLP.

1. Introduction

Natural language generation (Zock and Sabah, 1988; Dale, Mellish and Zock, 1990) forms an important component of many natural language applications, e.g., man-machine interface, automatic translation, text generation, *etc.* Bag generation (Brown, Cocke, *et al.*, 1990) is one of natural language generation methods. Given a sentence, we cut it up into words, place these words in a bag and try to recover the sentence from the bag. In corpus-based approach (Church and Mercer, 1993), a language model should be provided to measure the possible candidates. Markov Model (Kuhn and Mori, 1990) and word association model (Church and Hanks, 1990) are two famous models in language modeling. Markov Model has capabilities to keep the linear precedence relations in the context, so that it is useful to the application of bag generation. However, the parameters are tremendous in high degree Markov Model. Word association model can capture the long distance dependency relations in the context under the postulation that the window size is the length of sentence. Thus, it is useful to the applications such as lexical choice. This paper will propose an Approximate Markov Model, which has merits of these two models.

2. Approximate Markov Model

Let $S = <*, w_1, w_2, ..., w_m, *>$ be an arrangement in bag generation. The star symbol marks the beginning (w_0) and the ending (w_{m+1}) of the sentence. The probability of S in trigram Markov Model is measured as follows:

$$P(S)=P(<*, w_1, w_2, ..., w_m, *>).$$

$$\cong P(*)*P(w_1|*)*\prod_{i=0}^{m-1} P(w_{i+2}|w_i^{i+1})$$

$$= \frac{\prod_{i=0}^{m-1} P(w_i,w_{i+1},w_{i+2})}{\prod_{i=1}^{m-1} P(w_i,w_{i+1})}$$

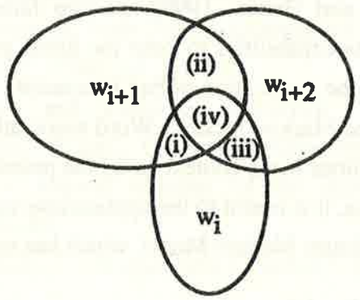
This formula utilizes trigram training table (numerator part) and bigram training table (denominator part) to compute the probability of an arrangement. It can be approximated by the following formula:

$$\frac{\prod_{i=0}^{m-1} \text{Min}(P(w_i,w_{i+1},1),P(w_{i+1},w_{i+2},1),P(w_i,w_{i+2},2))}{\prod_{i=1}^{m-1} P(w_i,w_{i+1},1)} \quad (1)$$

where Min denotes a minimal function,

$P(w_i,w_j,j-i)$ is the probability of a directed word pair (w_i,w_j) whose distance is j-i, e.g., $(w_i,w_{i+1},1)$ denotes w_i is followed by w_{i+1}.

By the notation of directed word pair with distance, the statement "w_{i+2} follows w_{i+1} and w_{i+1} follows w_i" (hereafter, $<w_i,w_{i+1},w_{i+2}>$) can be represented as $(w_i,w_{i+1},1)$, $(w_{i+1},w_{i+2},1)$ and $(w_i,w_{i+2},2)$. Consider the following figure. Assume parts (i), (ii) and (iii) correspond to the probabilities of $(w_i,w_{i+1},1)$, $(w_{i+1},w_{i+2},1)$ and $(w_i,w_{i+2},2)$, respectively. In this way, part (iv) denotes the probability of $<w_i,w_{i+1},w_{i+2}>$. From this figure, we know $P(w_i,w_{i+1},w_{i+2}) \le P(w_i,w_{i+1},1)$, $P(w_i,w_{i+1},w_{i+2}) \le P(w_{i+1},w_{i+2},1)$ and $P(w_i,w_{i+1},w_{i+2}) \le P(w_i,w_{i+2},2)$. Thus, the minimum of $P(w_i,w_{i+1},1)$, $P(w_{i+1},w_{i+2},1)$ and $P(w_i,w_{i+2},2)$ can be used to approximate $P(w_i,w_{i+1},w_{i+2})$.



The model formulated by (1) is called *Approximate trigram Markov Model*. Similarly, the following n-gram Markov Model:

$P(S)=P(<*, w_1, w_2, ..., w_m, *>)$

$$\cong P(*)*P(w_1|*)*P(w_2|*,w_1)*\ldots*P(w_{n-2}|w_0^{n-3})*\prod_{i=0}^{m-n+2} P(w_{i+n-1}|w_i^{i+n-2})$$

$$= \frac{\prod_{i=0}^{m-n+2} P(w_i^{i+n-1})}{\prod_{i=1}^{m-n+2} P(w_i^{i+n-2})}$$

can be approximated by:

$$\frac{\prod_{k=0}^{m-n+2} \text{Min}_{i,j(k\le i<j\le n+k-1)} P(w_i, w_j, j-i)}{\prod_{k=1}^{m-n+2} \text{Min}_{i,j(k\le i<j\le n+k-2)} P(w_i, w_j, j-i)} \quad (2)$$

Formula (2) denotes *Approximate n-gram Markov Model*. Assume the vocabulary size is V, and the average sentence length is L. The number of parameters of Approximate Markov Model is always $O((L-1)*V^2)$ no matter which degree it has. Markov bigram and trigram Model have $O(V^2)$ and $O(V^3)$ parameters, respectively. The number of parameters multiplies by V when the degree increases by one. Thus, Approximate Markov Model can be used to enlarge the window size, when the parameter issue is considered.

3. Bag Generation Algorithm

The bag generation algorithm under (Approximate) n-gram Markov Model is shown below.

```
insert starting node into queue
while not empty queue do
begin
  initialize an empty list
  repeat
    remove a node from queue, and assign it to current node
    if current node ≠ final node then
    begin
      expand current node and
      merge to the list if any two paths satisfy all of the following conditions:
      (1) the path length should be longer than n-1.
      (2) the lengths of these two paths should be equal.
      (3) the last n-1 nodes on these two paths should be equal.
      (4) these two paths should cover the same words.
    end
    else merge to the list
  until empty queue
  if current node ≠ final node then assign list to queue
end
generate the result from list, and check whether it is error or not.
```

The merge operation keeps the path with higher probability, and discards the path with lower probability. The four conditions in the above algorithm should be met if dynamic programming technique is used. The following proposition clarifies this point for Markov Model. Approximate Markov Model has the similar proof.

Proposition. The merge operation should obey the following conditions, if n-gram Markov Model is adopted:

(1) The path length should be longer than n-1.
(2) The lengths of these two paths should be equal.
(3) The last n-1 nodes on these two paths should be equal.
(4) These two paths should cover the same words.

Proof:

The first two are the basic definitions for n-gram Markov Model. In this model, the system will use the last n-1 words to predict the probability of the current word. Let the probabilities of two paths H_1 and H_2 be $P(H_1)$ and $P(H_2)$, and $P(H_1) > P(H_2)$. When the next word w_m ($m \geq n-1$) is read, their probabilities become:

$P(H_1) * P(w_m \mid w_{1(m-n+1)}, ..., w_{1(m-1)})$ and

$P(H_2) * P(w_m \mid w_{2(m-n+1)}, ..., w_{2(m-1)})$, respectively.

If the last n-1 words are the same, i.e., $w_{1(m-n+1)}=w_{2(m-n+1)}, ..., w_{1(m-1)}=w_{2(m-1)}$, then the former is still larger than the later. However, if the last n-1 words on these two paths are not the same, then the former may be smaller than the latter. Thus, merging may introduce the error results.

In fact, the first three conditions are enough for the other Markov-based applications such as phone-to-text transcription, *etc*. However, there is a problem in bag generation application, if we do not obey the last condition either. Consider a general case. Let the two paths H_1 and H_2 have the following forms:

H_1: $w_{10}, w_{11}, ..., w_{1(m-n)}, w_{(m-n+1)}, ..., w_{(m-1)}$ and

H_2: $w_{20}, w_{21}, ..., w_{2(m-n)}, w_{(m-n+1)}, ..., w_{(m-1)}$.

If $\{w_{10}, w_{11}, ..., w_{1(m-n)}\}$ is not equal to $\{w_{20}, w_{21}, ..., w_{2(m-n)}\}$, there must exist some w_{1i} and w_{2j} such that $w_{1i} \neq w_{2j}$. If $P(H_1) > P(H_2)$, then the path involving w_{1i}, i.e.,

$w_{20}, w_{21}, ..., w_{2(m-n)}, w_{(m-n+1)}, ..., w_{(m-1)}\ w_{1i}$

will be neglected. This path may have higher probability, so that error occurs. ■

The cost paid by the Approximate n-gram Markov Model is: each minimal value in the numerator part and denominator part of Formula (2) is derived from n*(n-1)/2 pairs and (n-1)*(n-2)/2 pairs, respectively. Consider the numerator part. For each tuple $<w_k, w_{k+1}, ..., w_{k+n-1}>$ ($0 \leq k \leq m-n+2$), its probability is determined by $P(w_i, w_j, j-i)$ ($k \leq i < j \leq n+k-1$). The complexity of an algorithm to select the minimum from n*(n-1)/2 pairs is $O(n^2)$. It is a terrible overhead. Here, a special data structure, i.e., a ring of n-1 elements, is adopted. Each element records the minimum of k+n-1-i probabilities $P(w_i, w_{i+p}, p)$ ($1 \leq p \leq (n-1)-(i-k)$). The index i is ranged from k to n+k-2. The minimum of the n*(n-1)/2 pairs can be computed from these n-1 elements. When k is increased by one, i.e., the tuple $<w_{k+1}, w_{k+2}, ..., w_{k+n}>$ is inspected, only these (n-1) elements are considered instead of n*(n-1)/2 pairs. In other words, the position in the ring for $P(w_k, w_{k+p}, p)$ ($1 \leq p \leq n-1$) is free, and is used to record $P(w_{k+n-1}, w_{k+n}, 1)$. $P(w_{k+p}, w_{k+n}, n-p)$ ($1 \leq p \leq n-2$) are compared with the corresponding elements in the ring. This can be done in O(n) time.

4. Experimental Results

BDC corpus, which is a Chinese segmented corpus, is adopted as the source of the training data. It includes 7010 sentences about 50000 words. For each sentence $S = <*, w_1, w_2, ..., w_m, *>$ in the training corpus, total (m+1)*(m+2)/2 directed word association pairs, which are of the form $(w_i, w_j, j-i)$ (where $0 \leq i < j \leq m+1$), are generated. The experimental results (distribution of error sentences) of bag generation by using Markov Model and Approximate Markov Model are shown in the following table. M*i* and AM*i* denote *i*-gram Markov Model and Approximate Markov Model, respectively.

sentence length	total test sentences	Markov Model				Approximate Markov Model				
		M2	M3	M4	M5	AM2	AM3	AM4	AM5	AM*n*
1	6	0	0	0	0	0	0	0	0	0
2	34	0	0	0	0	0	0	0	0	0
3	121	0	0	0	0	0	0	0	0	0
4	213	1	0	0	0	1	0	0	0	0
5	297	0	0	0	0	0	0	0	0	0
6	329	3	0	0	0	3	0	0	0	0
7	234	4	0	0	0	4	2	1	0	0
8	216	11	0	0	0	11	1	1	0	0
9	183	6	0	0	0	6	0	0	0	0
10	170	8	0	0	0	8	0	0	0	0
11	129	11	0	0	0	11	0	0	0	0
12	68	13	0	0	0	13	1	1	0	0
total	2000	57	0	0	0	57	4	3	0	0

It is trivial that AM2 is equal to M2. The other results demonstrate that the power of approximate Markov Model is close to that of Markov Model.

5. Concluding Remarks

This paper proposes a directed word association model with distance to approximate Markov Model. It can increase the degree of language model, and keep the number of parameters unchanged. The experimental results show that the performance of Approximate Markov Model and Markov Model is very close. Besides, the training knowledge for bag generation can be also applied to lexical selection. The co-occurrence of a word pair can be computed easily by sum of the related directed word association pairs. The uniform knowledge facilitates statistics-based machine translation design.

References

Brown, P.; Cocke, J., *et al.* (1990) "A Statistical Approach to Machine Translation," *Computational Linguistics*, 16(2), 79-85.

Church, K.W. and Hanks, P. (1990) "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, 16(1), 22-29.

Church, K.W. and Mercer, R.L. (1993) "Introduction to the Special Issue on Computational Linguistics Using Large Corpora," *Computational Linguistics*, 19(1), 1-24.

Dale, R.; Mellish, C. and Zock, M. (eds.) (1990) *Current Research in Natural Language Generation*, Academic Press, London, England.

Kuhn, R. and Mori, R.D. (1990) "A Cache-Based Natural Language Model for Speech Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6), 570-582.

Zock, M. and Sabah, G. (eds.) (1988) *Advances in Natural Language Generation: An Interdisciplinary Perspective*, Printer Publishers, London, England.

"Sud'ba" and "Rok". Semantics of Noun As An Object Of Conceptual And Quantitative Analysis

Ljudmila O. Cherneyko
Moscow State University
Philological Faculty
Russia. 121433, Moscow, Malaya Filevskaya, 38-51,
Phone: (095) 146-2523

Vladimir A. Dolinsky
"Lingua" Cooperative
Russia. 129345, Moscow, Ostashkovskaya,9-2-98,
Phone: (095) 475-8384.
E-mail: polikarp@logos.msu.su

Topical paper

AREA: Psycholinguistics, semantics, epistemology, methodology

Summary:
Semantic analysis of Russian words "sud'ba" and "rok" ("fate" and "lot") based on text material, vocabulary defi nitions and distribution of responses to these words as stimuli in word association test

The understanding of fate as a super power affecting both a human being and the humanity at large is invariant to the lexicographical interpretations of the word "sud'ba", the latter do not coincide and are inadequately differentiated. It is "inevitability in earthly existence" (Vladimir I. Dal'). The attempts to formalize the common notions of Russian speakers into dictionary definitions reveal that dictionary makers' approaches fail to coincide. .Rational comprehending of meaning associated to a definite word rests on an image, an idea, a Gestalt rather than on empirical knowledge. Characteristics comprising a concept's significate, an intensional are but derivatives of traditional notions born in the Russian language and cultural element.

The proto-name "sud'ba" does exist in the real world, though it is an ideal world, not a physical one. Man defines "sud'ba" as something he himself created, i.e. the idea implementing his actual dependence on the world, a dependence endowed with supernatural powers.

The part-of-speach characteristics of the word "sud'ba" testifies to the underslying phenomenon as a substance, i.e. something that exists in itself as bearer of a quality, a state, a process. A substantive denotes "a certain sum total of features, none of which is named" (A.Peshkovsky). Though the word still retains its inner form, this feature (sud-it') is beyond rational understanding and this phenomenon is perceived irrationally, by intuition, by feeling rather than thinking. That's why componential analysis - a ra-tio instrument in preparing the concept significate gives way to conceptual analysis, an instrument taking into account that which is perceived by intuition. The concept comprises the notion but does not exhaust it, taking in the whole meaning of the word - both denotative and connotative, on the basis of an idea. Ideas as such can be devided into ideas - recollections (images of things perceived earlier) and ideas - imaginations (fantasy images), mythologemes, as it were.

"Sud'ba" is one of such mythologemes. The specific character of notions consists in their associative nature, they "can call forth one another" (A.Peshkovsky). It is not associations that arise from the word combinability (secondary associations) but the word combinability is conditioned by its associative potential. The combinability of a word is indeed the exterior, surface manifestation of deeper associative contours of a name. The symbolization of an abstract essence underlying an abstract name realises itself through combinability. The concept i.e. the generalised image of a word is made up of several Gestalten that vary in different languages as there vary the world pictures of their speakers just like pictures in children's kaleidoscope. The problem is "to bring to the light of the scientific reasoning the things which have shaped themselves and exist in the objective psychological world apart from any science" (I.Boduin de Courteneau), that "field of inner experience which is segmented differently by the cold analitical reason and creative imagination of language-makers" (W.von Gumboldt).

"Sud'ba" (fate) "present itself to us in various appearances" (M.Montaigne) which are hardly possible for Russian language consciousness to be calculated though allow to reveal "auxiliary subjects" of implicit metaphor in texts.

1. The personification of "sud'ba" [S] (fate). S.- a person. An old woman - s. Whims of s. To make s. angry. S. protects, guards. (Social status) S.- a mistress, queen, patroness. Gifts of s. To thank s. Power of s. A slame, victim of s. A strike of s. (Partner in a game) S. has made a move. His play with s. (Resisting) To fight s. To curse s. S. chases, pursues, lies in wait.

2. Sud'ba as a text, a source of information. It is written on his s. To seal s. To repeat s. A sign of s. To believe in s.

3. Forms of existence of "sud'ba". Point. Circle. Line. Spiral. Whole, consisting of parts. Part of (whole). Time. Elements. Mortal creature. Material object (to hand over, to leave, to breake, to be master of s.). Artifact (to models., creator of one's own s.). Value.

4. Substantiating (objectivising) sud'ba. Thread. Road. Land (to plough s.). Premises, dwelling (to break into s., to put one's s. in order = to tidy one's s.). Clock. Vessel. Clothing. Animal (to take s. by the horns). Tree branch (to break off s.). Money (to waste s., to pay off s.).

In texts of poetry "sud'ba" is all-embracing, bitter, miserly, menacing, it wearies, poisons (destroys), torments, subdues (A.Fet), a cruel wizard, it punishes, temptible

(A.Blok), it breathes, a miser, there are gaps in it, it is shaped (B.Pasternak), a channel, it is digging smth (I.Guberman), it encroaches on, it is a disturber, a polyglot, it burns out, it's splits up, it's a book, a game, an admixture of geography to time (I.Brodsky).

The dictionary definition of "sud'ba" given by S.Averintsev reads: "In mythology, in irrationalist philosophical systems as well as in common consciousness it is unreasonable and unfathomable predetermination of events and human actions". The idea of "sud'ba" is opposed to both scientific (rational) concept of cousability and religious notion of predetermination (teleologic determination), for whom "sud'ba" is a pseudo-name of God, according to this "it is not God who hides from man, it is man who conceals his faith in God from himself" (Alexis II). In V.Dal's dictionary "sud'ba" and "rok" [R] are distinctly opposed: "sud'ba" is the one who chases, arrests, passes sentences on, while "rok" is the executioner, hangman. R. is inevitable, he does not take decisions, he implements them. One can't avoid s. S. ties one's hands and r. seeks one's head. One is destroyed, depressed by r. R. gets the guilty. S. is associated with destiny, lot, fatum, life. The opposition following the axis freedom - no-freedom.

In the psychic world of a Russian-speak person a unique picture (idea) of "sud'ba" has been established. The conceptual analysis allows to only supply with the material of texts and dictionaries in order to estimate the share of one or another Geshtalt in the whole picture of a language notion (image), but it does not provide a complete answer. The statistically reliable answer which is topical for contemporaries can be only achieved by using the results of proceedings of word association tests.

Quantitative analysis of associative fields serves as a means of revealing objective semantic characteristics of linguistic units and is indispensable source of information on language-cultural Gestalten, reflected in images of word sense.

The words "sud'ba" and "rok" were given as stimuli to two different groups of tested Moscow students in 1992-1993. The parameters of distribution (obtained responses) and the list of associations (with frequences not less than F = 2) are given in the table (1).

The semantic connection between "sud'ba" and "rok" is clearly indicated by the fact that the responses to the one by the other as an association takes the third and first place in the hierarchy (S.: F_3(R.) = 28, R.: F_1(S.) = 96).
By excluding from the associative field of "rok" the reactions caused by its homonym (engl. "Rock", not marked with asterisk *), we can build the hierarchy of associations common to the both stimuli. The ratio (A) of differences and sums of frequences of identical responses (i) points to the degree of specificity (A=1) or community (A=0) of a given association in the fields of the both stimuli.

$$A = \frac{F_i(s.) - F_i(r.)}{F_i(s.) + F_i(r.)}$$

To the common associations there are referred: stars, life, death, cruel, hard, blind, fatum, himbleness. The spesificity of "sud'ba" is in: man, uncertainty, happy, God, line, way, hand, cross. The spesificity of "rok" is in: chase, plague, terrible, fear. "Sud'ba" is associated with obedience, while "rok" - with indifference and confusion, and they both are associated with humbleness. The archetypes of "sud'ba" angel, scales, road, cross, ring.

While pressure, chastising sword, tears, plague are the archetypes of "rok".

Table (1)

"Sud'ba" and "rok" associative fields

FATE (sud'ba)	S = 439, N = 393, m = 119, l	"-" = 46, L = 172, F = 35 l	LOT: (rok)	S = 440, N = 417, m = 94, l	"-" = 23, L = 125, F = 96. l
35	c[illegible]an	important	*	96	fate (sud'ba)
34	life	scales		51	music
28	lot	time		35	rock-n-roll
14	inevitability	road (doroga)		(*)22	hard
12	villainess	only [adj.]		20	lesson (urok)
8	fatum	cruel (zlaya)		9	guitar
7	uncertainty	outcome (ishod)	*		cruel (zloy)
	happy	ring		7	musician
6	carma	kopeck		6	metal
	hard	a Fate (Moira)		5	jazz
5	such	obedience			performance
4	God	predictination		4	Alice
	cruel (zhesto-kaya)	wonderful desidedness			Beatles cornucopia
	turkey	probability	*		fatum
	line	(sluchaynost')			noise
	my	mystery		3	rattle
	unavoidable	each has his own			DDT
	way (put')	fatalist	*		life
	hand	fortune	(*)		star
	death	chiromancy			punk
	fatality	good	*		chases
3	stars	l	*		death
	cross				and Roll
	there is no			2	studio
	fatal (rokovaya		*		inconsolability
	happiness				loudly
	painful				group
	fatalism				tape-recorder
	fatal		*		inescapable
2	angel				noisy
	future				

Note: S - number of subjects; "-" - number of zero reactions; N - number of received responses; L - assortment of associations, or A-glossary; m_1 - number of associations of frequency 1; F - frequency of association ranked 1

The associative field of "rok" is absorbed or makes part of the field of "sud'ba" and this fact is testified by the frequency spectrum of the total distributions for the two stimuli.

"The method of dividing the field of thought by means of language variability (diversity) has been little tested as yet, but it does not become less possible or important because of it. However rich and fruitful a language might be, it is never possible to imagine the real sense, the sum total of all integrated characteristics of a word, denoting a non-physical object, as the definite and final value" (W. von Gumboldt).

"Судьба" и "Рок". Семантика существительных как объект концептуального и квантитативного анализа

Л.О. Чернейко,
В.А. Долинский

Резюме:

Семантический анализ русских слов " судьба " и " рок ", основанный на текстовом материале, словарные определения и распределение реакций на эти слова как стимулы в словарно-ассоциативном эксперименте.

Multicomponent Names of International Organizations as Terminological Units

L.A.Chizhova, M.V.Shibaeva,
Lomonosov Moscow State University,
Philological Faculty,Chair of General and Comparative Linguistics
Phone: 315-30-18; 417-51-91.

Topical paper

AREA: Description of some aspects of organization of terminological units.

SUMMARY:
The description of translation and structure of multicomponent names of international organizations as interlevel and most productive nominational units of terminological character, statistic analysis of their formal and semantic structure.

In this report the multicomponent names of international non-governmental organizations in the consultative status with the Economic and Social Council (UN) which are registered in UN Terminology bulletin No.331 are considered. International organizations are associations of social organizations, private companies and private persons of different countries which are organized to achieve common political, economic, social and cultural gains. For example:
Association of West European Shipbuilders AWES -
Ассоциация западноевропейских судостроителей АЗЕС,
International Council for Philosophy and Humanistic -
Международный совет по философии и гуманитарным наукам.

The statistic and qualitative analysis of formal semantic structures of the English names and their Russian translations is carried out. We have come to the conclusion that English and Russian names differ widely in the structures itself, in the number of components and in the nature of their semantic relations because the Russian language expresses the meaning more specifically and in detail than the English language. For example:

Asian Cultural Forum on Development	Азиатский форум по культурному развитию
N	N
/ | \	/ |
A A on N	A поN - A

In the considered multicomponent terminological combinations we single out such elements which are the basis for the whole name because they bear the general information load. These elements can be called basic or kernel.

The basic elements of these combinations are not homogeneous by origin and semantic importance. Some of the basic elements designate the general notions which are typical only for ergonims (union - союз, council - совет). Some other elements are taken from different semantic fields and are the structural and semantic support of these combinations (project - проект, experiment - эксперимент). Some basic elements are stylistically neutral (organization - организация), but others are taken from the elevated style (fellowship - братство, league - лига).

The statistic analysis of translation and position of a basic element in the multicomponent names of organisations is carried out. It is turned out that the most regular position of a basic element is inposition although the general tendency for the English terminology is postposition of a basic element. This peculiarity of multicomponent names of organizations can be explained by their international nature and it is always reflected in the name (preposition of the word "international"):
International Federation for Human Rights -
Международная федерация прав человека.

The basic elements are usually accompanied by terms, anthroponyms, toponyms which indicate the concrete sphere of activity, location and the creator of the organization. These words are specially organized in a multicomponent name and in most cases change some of their categorial features under the effect of its multicomponent structure. Anthroponyms contain the "intellectual" information which can be taken from reference books and dictionaries. Toponyms have differential and address functions. They either help to distinguish one organisation from another:
Association of African Universities -
Ассоциация африканских университетов,
or just indicate its location:
Central Bureau of Statistics (Kenya) -
Центральное бюро статистики.

In conclusion it should be pointed out that a multicomponent name of organization is an integral nominative unit which is characterised by the indissoluble connection of its components. It has one denotative meaning which is conveyed by the whole name. That is why the multicomponent terminological combinations are considered to be interlevel and the most productive nominational units of international organizations.

Многокомпонентные названия международных организаций как терминологических единиц

Л. Чижова
М. Шибаева

Резюме:
Описание перевода и структура многокомпонентных названий международных организаций как межуровневые и наиболее продуктивные номинативные единицы терминологического характера, статистический анализ их номинативной и семантической структуры.

MODELS OF BILINGUAL MEASUREMENT AND THEIR ADAPTABILITY IN THE INDIAN CONTEXT

Amitav Choudhry
Indian Statistical Institute
Calcutta, India
E-mail: chou@isical.ernet.in

ABSTRACT

With a view to develop tests to measure bilingual ability in the Indian context I have examined some existing language tests and tried to analyse their validity and adaptability in the Indian situation. The tests may be divided into two categories; [1] Discrete Point Language tests for bilinguals and [2] Discrete Point and Integrative Oral Language tests or Quasi Integrative approaches. I have tried to develop Methodological concepts supported by empirical studies to assess the reliability of available tests on measurement of bilingual ability in the Indian context.

INTRODUCTION:QUANTIFICATION OF BILINGUAL ABILITY:

In the realm of evaluation of bilingual children are divergent views, definitions and theories on how children acquire first, second and subsequent languages. Various aspects of verbal expressions of children can serve as indicators of the level of maturity in language development. Verbal expression can often be the medium through which social interaction, cognition and other linguistic behaviour is examined and proper comprehension of the probable verbal performance of children can often be of great value in determining the most appropriate design for an empirical study, therefore a sociological orientation to both first and second language has to develop. Researchers have become aware of and concerned with the importance of social setting, interactors, and topic of discourse.[cf. Cazden 1970, 1972a, 1972b, especially with reference to minority dialects and bilingualism. cf. Fishman, 1972 and Labov,1972.]

Language tests developed before the Chomskyan revolution naturally neglected the thinking that language was a series of seperate or discrete points which when added up, made the whole. Language was not viewed as a synergistic and social phenomenon (Erickson, 1981). Some tests measured several discrete points, for e.g. The Michigan Picture Language Inventory (Lerea, 1958). Most of the discrete point tests also have limited or questionable statistical support, especially in regard to their use with minority children.

THE NEED TO MEASURE BILINGUAL ABILITIES:

In order to conduct meaningful 'dynamic' studies of social relations, personality growth and the like, we must first have normative studies of childrens' developmental performance in articulation, vocabulary, and communication of meaning, in the languages they are exposed to and subsequently acquire. Way back in 1930, Mc Carthy and Davis (1937) established a ground work of such studies, delineated areas of investigation and supplied norms of fundamental significance. Templin [1957] improved on these efforts on two counts, (i) He collected in one study in one sample of children, normative measures of articulation of speech sounds, sound discrimination, sentence structure, and vocabulary, thus allowing the study of interrelationships among these measures. (ii) The design permits a comparison of contemporary norms with data established on similarly selected children in an earlier period. But today these studies, especially reliance on comparative studies do not have much relevance in the Indian bilingual context, because one cannot expect the exact situations to exist, say even five years from now. The children tested today in a particular linguistic setting with the help of certain control variables like social status, bilingual status, academic achievement level, linguistic environment of the child, etc. and also the associative linguistic experience proportionate to the changing times may not remain exactly the same to enable one to draw one to one inferences or for that matter even comparative inferences whereby one can establish a norm emphazising the pattern of change.

In this paper I have attempted to analyse the validity (in the Indian context), of a few tests used to assess first and second language dominance. The tests may be divided into two categories, (i) Discrete Point Language tests of Bilinguals and (ii) Discrete Point and Integrative Oral Language tests or Quasi Integrative approaches.

DISCRETE POINT LANGUAGE TESTS:

These are currently in use in most academic second language and bilingual instructional settings. Language was seen as a series of distinct structural units (eg. phoneme, morphemes), and mastery of each of these seperate units was judged to be equivalent to mastery of the language. Adequate models to test each of these individual structural units were designed;therefore discrete point tests were developed.

QUASI INTEGRATIVE APPROACH:

A combination of discrete point and integrative oral language production instruments are currently being used to assess bilingual children. These approaches to assessment are attempts that recognize the importance of spontaneous language sampling as the basis of assessment. Examples of these approaches are the Oral Language Evaluation (OLE) Silvaroli & Maynes (1975), Basic Inventory of Natural Language (BINL) Herbert (1977), and Bilingual Syntax Measure [BSM] Burt et al (1976). Each of these approaches call for a sample of natural language, cued by pictures, scored in a discrete point fashion, with emphasis of syntax, vocabulary and length of response.

Experiments on three models were conducted by me to determine their effectiveness and adaptability in the Indian context. The subjects were pre-school and primary school age children (4+, 5+ and 6+) with Bengali as L1.

BILINGUALISM AMONG PRE-SCHOOL CHILDREN:A CASE STUDY

Firstly I conducted an experiment to test the effectiveness of the Mc Carthy (1930) and Davis (1937) method to test the various aspects of the verbal utterances among pre-school age bilingual children. 25 verbalisations each were obtained from 30 subjects in the age range 2 - 4 years to determine the degree of bilingualism in the pre-school age child whose mother tongue is Bengali. The languages tested were Bengali and Telugu. Table:1 shows the performance of the children in the two languages concerned:

Age in years	Total	Mean Bengali	S.D.	#	Total	Mean Telugu	S.D.
2.0	16.50	2.75	0.5244	#	11.00	1.83	0.5177
2.5	18.00	3.00	0.3162	#	15.50	2.58	0.3768
3.0	21.75	3.63	0.2098	#	17.50	2.92	0.3768
3.5	25.50	4.25	0.5235	#	18.50	3.08	0.2049
4.0	32.50	5.42	0.5621	#	24.75	4.13	0.4669

Table-1: Performance of children over 5 age groups in Bengali and Telugu to determine the number of words uttered in 25 verbalisations each. [N = 30]

Age in years	't'-value	level of significance
2.0	3.05	.05
2.5	2.0752	---
3.0	4.0221	.01
3.5	5.0837	.001 [high]
4.0	4.3298	.01

Table-2: 't' values of the two languages, Bengali and Telugu over 5 age groups, for boys and girls

Results show that since the calculated value is greater than the theoretical value even at 10% for the age group [3.5 yrs] and at 1% for the age group [3.0yrs.] and [4.0yrs.] and at 5% for the age group [2.0yrs.] we can conclude that except for the age group [2.5yrs.] the scores or the diffeence in the number of verbalisations for the two languages namely Bengali and Telugu are significant. Therefore in accordance to Titone's [1972] prediction traces of bilingualism were found even among 2 year olds as far as verbalisations in the two languages are concerned though the number of words per verbalisation in Bengali is better than the number of words per verbalisation in Telugu among pre-school-age Bengali children.

JAMES LANGUAGE DOMINANCE TEST (James 1975):

The test was carried out on 30 kindergarten children whose L1 was Bengali spread over 5 age groups, i.e., 3 years, 3.5 years, 4 years, 4.5 years and 5 years, with 3 boys and 3 girls in each age group. The test items included 20 pictures which would evoke one word or two word responses. The pictures included items covering all spheres of daily life of children, from household articles or items, celestial objects, animals to objects of nature, etc. Test administration time was of 20 minutes duration. Each correct response was awarded 2 points, and a minus point was deducted for a half correct response where 2 word responses were expected, and a proper grammatical sequence was required. Questions asked were mainly in the form of, "What is this?", Who is this?" or "Where is the -?". Phonological variations were overlooked if it was only one per word and for more than one, one minus mark was awarded. The maximum score possible per subject was 40 points.

Based on the results, the subjects were put into 3 categories to ascertain their language dominance, bilingual proficiency and also whether they were bilinguals with dominance in one particular language or whether they were proportionate bilinguals. The languages in question were Bengali [L1], English [L2] and Telugu [L3].

Categories - A = L1 dominant
B = Bilingual plus L1/L2/L3
C = Proportionate bilingual [L1/L2] or [L1/L3]

Age Groups [Years]	A L1	B [L1	L2	L3]	C [L1 / L2]		[L1 / L3]	
3.0	46.52	24.50	24.06	4.81				
3.5	29.75	21.45	11.71	.98	19.02	17.08		
4.0		19.44	12.96	.93	32.41	34.25		
4.5		25.55	19.38	3.52	16.30	18.50	8.8	7.93
5.0		12.66	35.80	1.31	24.45	25.76		
Freq. dist. in %	13.91	20.58	20.96	2.25	18.98	19.74	1.88	1.69

Table-3: Frequency distribution in [%] of the 3 categories given seperately for each of the 5 age groups.

Results show that they were very few phonological variations or incorrect grammatical sequences that would adversely

effect the scores of the subjects. L1 dominant subjects were found only amongst 3 year and 3.5 year olds. Regarding bilingual children except for 5 year olds there were more L1. dominant than L2 dominant subjects amongst the other age groups. The % of L3 dominant bilinguals was only 2.25%. Except for 3 yr olds, proportionate bilinguals were found for all other age groups. i.e. [L1/L2]. There were a large number of proportionate bilinguals amongst 4 year olds. L1/L3 proportionate bilinguals were found only for 4.5 year olds. Scores improved with age, but there was no perfect score obtained by any of the children. This reflects the fact that responses to the stimuli was best amongst 4.5 and 5 year olds. They exhibit an adequate vocabulary in both L1 and L2 and most children maintained a proper grammatical sequence for two word responses. There were also a limited number of cases of interference or mixed responses but these did not have any recognisable impact on the scores. Most of the children were found to be language proficient in both L1 and L2 and not proficient in L3.

Finally this test proved that as far as vocabulary is concerned it is not necessary that children should be dominant only in L1. The bilingual children can also be dominant in L2. As children are exposed to the school environment the number of proportionate bilinguals increase. Though most children were also exposed to L3 there were no L3 dominant cases. The home environment language was either L1 or L2.

THE BILINGUAL SYNTAX MEASURE (Burt et al 1981)

The test was carried out on 12 school age children, both boys and girls in the age group of 4+, 5+ and 6+, whose L1 was Bengali, L2 English and L3 Telugu. All the children were from English medium schools. They also had spent all their lives in Hyderabad and Secunderabad.[India]. Each version, (to test L1, L2 & L3 seperately) consisted of 20 questions, not necessarily translation equivalents that were intended to elicit particular grammatical structures about a series of 7 pictures which were self expresive. Responses were recorded seperately in 3 booklets and later analysed for acceptability and point value. The rules for evaluation were the same as used for the original test as cited earlier. Apart from this a Global test was applied keeping the criteria in mind to assess the degree of acceptability on a 6-point rating scale given below:

Acceptable [+]			Unacceptable [-]		
1	2	3	4	5	6

The results of the two tests are seperately tabulated below:

95%	to 100%	Proficient	[P]
85%	to 94%	Intermediate	[I]
45%	to 84%	Survival	[S]

Age	(LANGUAGE) Bengali[L1]	English[L2]	Telugu[L3]
4 years	60.41 [S]	61.58 [I]	82.70 [S]
5 years	71.29 [S]	65.75 [I]	89.37 [S]
6 years	71.37 [S]	64.37 [I]	95.20 [S]

Table-4: Percentages of Competency levels of 12 children [with Bengali as their L1] in L1, L2 and L3

Age	(LANGUAGE) Bengali[L1]	English[L2]	Telugu[L3]
4 years	2.00	1.67	3.35
5 years	1.43	1.57	2.82
6 years	1.20	1.30	2.12

Table-5: Average Ratings of 12 children [with Bengali as their L1] in L1, L2 and L3

As seen in the analysis of the other tests the results varied with age. Though the percentage of proficiency improved with age, subjects tested for competence in L1 and L3 fell into the 'Survival' category but when tested for L2 they did better and 4+ and 5+ subjects were put in the 'Intermediate' category and 6+ subjects were put in the 'Proficient' category. At the same time subjects were more competent in L1 than in L3. The reason for the improved proficiency in English was that from 4+ onwards there is a marked exposure to English at the school level, with a diminishing scope of interaction in both Bengali and Telugu.

The results of the Global Test are in interesting contrast to the modified version of the test as used by Burt et al (1981). Though the average ratings are slightly better when tested for proficiency in L2 the degree of accept-

ability of responses of 5+ and 6+ subjects is better in L1 when compared to the acceptability ratings of 5+ and 6+ subjects in L2. Therefore te difference in using the Discrete Point test and the Global Test are quite evident. The Global Test is a more careful evaluator of natural language samples than the Discrete Point Test.

There is need for further research and more indepth longtitudinal studies of child language acquisition in multilingual settings. Whatever models have been used or may be used must also consider pragmatic and ethnological techniques. Also needed is an accurate determination of what language a bilingual child uses in which situation and what is needed for effective communication within the situation that the child encounters in daily life. Such studies could serve as a basis for more appropriate and realistic assessment of a bilingual child's language proficiency. According to Day (1981) an awareness that language is more than just a sum of its discrete parts stimulated the development of integrative tests of language proficiency which would illuminate to a larger extent the subjects' underlying total competence rather than use tests which only measured awareness of the various units of language. Global measures would in the long run be more predictive of a person's actual performance in a second language than the previously used discrete point tests. And finally the need of the hour is to develop tests which can tell you explicitly what a child can do and not what a child cannot do.

****

Topical Paper: (Measurement of Bilingualism)

fuzzy logic. As opposed to existing models our model takes into account natural language reality and natural language restrictions. As opposed to time logic, tenses are only secondarily taken into account, and seen as only one, though important and (proto)typical, means of indicating 'time'. Not only tense, which has received most attention in the literature, but also temporal prepositions, adverbs and open class lexical items, especially nouns, as well as word order, amongst others, determine temporal reference. In this paper we focus on lexical time expressions. Lexical time indications can be subdivided into several ways. For the remainder it may suffice to list the following semantic division and characterization of time expressions. Relational or relative expressions indicate a point or interval that relates to a given time fact (e.g. *'around 10 a.m.*). They are generally more vague than situational indicators, that point to a time fact itself (e.g. *in May*). Relational expressions refer to a relation with a time point or interval and this relation is an anterior, posterior or approximative one. Unbound or free expressions do not refer to past, present or future (e.g. *at two o'clock*), while bound expressions do (e.g. *in May 1944*). Both relational or situational, and bound or unbound time expressions may indicate a time point or a time interval.

Time expressions often show some vagueness in degree, as opposed to vagueness in criteria. The first kind of vagueness is to be found in, for instance, *a tall man* or *an old man*, where the vagueness resides in the fact that one and only one well-determined criterion is being scaled, in these cases "length" and "age" respectively. Vagueness in criteria, on the other hand, can be found in expressions like *an intelligent man* or single items like *religion* and *art*: most often different criteria are called upon in naming something *religion* or *art*. Hence, this kind of vagueness is multidimensional. It should be noted that many lexemes are vague in both senses (e.g. *a big house* (vague in criteria) → *a big house* (i.e. with many bedrooms; vague in degree)). As time is a one-dimensional fact, only vagueness in degree and not vagueness in criteria is involved (and vagueness in degree itself is a matter of degree, as relational expressions are often more vague than situational ones). Moreover, time can be expressed numerically (which makes time objectifiable). All this should facilitate a formal representation of vague time expressions by means of fuzzy set theory (Devos, 1994).

Lexical vagueness can be found in three major classes of time expressions: especially in (1) approximative time indications (e.g. *around 6 p.m., around 1972*), but also in (2) indicators of half closed (or half open) intervals, i.e. indicators of posterior and anterior relations (e.g. *shortly before 6 p.m., some time after the holidays*), and in (3) indicators of frequency (e.g. *often, seldom, sometimes*). In the next section we will focus on classes (1) and (2), as (3) is definitely not primarily a class of time expressions: indicators of frequency only mention the frequency of an action or event, though this is done against some temporal background of course.

3. Modelling Periods of Time by Means of Fuzzy Set Theory

In the next paragraphs we will discuss the different ways of representing periods of time one by one. In each case some examples are given.

3.1. Vague time intervals

A *fuzzy interval* is a normalized convex fuzzy set in $\mathbf{R}$, i.e. the membership function μ satisfies the following:

$$\exists\, x \in \mathbf{R} : \mu(x) = 1$$

$$\forall\, x_1, x_2 \in \mathbf{R}, \forall\, \lambda \in [0,1] : \mu(\lambda x_1 + (1-\lambda) x_2) \geq \min(\mu(x_1), \mu(x_2))$$

A *fuzzy closed interval* is an uppersemicontinuous fuzzy interval, i.e. the membership function μ satisfies in addition:

$$\forall\, y \in \mathbf{R}, \forall\, x \in \mathbf{R}, \forall\, \epsilon > 0, \exists\, \delta > 0 : \mid x-y \mid < \delta \Rightarrow \mu(x) < \mu(y) + \epsilon$$

which expresses that in each element the degree of membership equals the maximum of the left and right limit value of the membership function in this element. The frequently used trapezoidal fuzzy sets are special cases of fuzzy closed intervals (Fig.1).

Such a fuzzy (closed) interval can be used to model vaguely expressed periods of time. A fuzzy time interval is an immediate fuzzy extension of the crisp notion of time interval: whereas a point of time x either does or does not belong to a crisp time interval, it can "belong" to the fuzzy time interval, modelled by the membership function μ, with a degree $\mu(x)$..

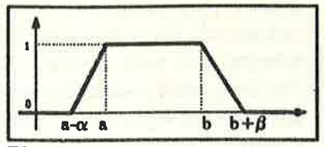


Fig. 1

All approximative time indications, indications that render a time interval or a time point approximatively, are represented in this model (e.g. *around 4 p.m., around 1972, about noon*).
This representation of periods of time has the drawback that no complete information can be derived concerning the starting point or the end point of the time interval (e.g. does 4.25 p.m. still belong to the interval *around 4 p.m.*?).

3.2. Starting point and end point of time

The (vague) starting point of time S as well as the (vague) end point of time E of the time interval can be indicated by using a fuzzy set (disjunctive interpretation, i.e. a possibility distribution. Two fuzzy time intervals can be deduced from this (Fig.2). The convex hull of both fuzzy sets S and E renders all time points that possibly belong (with a certain degree) to the interval. The intersection of the fuzzy set of time points that necessarily come after S, and the fuzzy set of time points that necessarily come before E, indicates all time points that necessarily (with a certain degree) belong to the time interval.

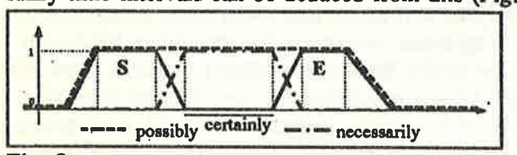


Fig. 2

Two crisp intervals over time can also be associated with S and E: on the one hand there are those time points that certainly belong to the interval, those certainly coming after S and certainly before E; on the other hand, the time points that certainly do not come after S and certainly not before E, definitely do not belong to the time interval.
Some examples of intervals expressed by two vague lexical items are *between midnight and dawn, between Renaissance and the Romantic Age,* and *between the beginning of next week and the end of the month.*

Remark:
The fuzzy set A1 of time points that *possibly* come after a vague time point A, is given by the membership function:

$$\mu_{A1}(t) = \sup_{s \leq t} \mu_A(s).$$

The fuzzy set A2 of time points that *necessarily* come after A, is given by the membership function:

$$\mu_{A2}(t) = \inf_{s \geq t} (1 - \mu_A(s)).$$

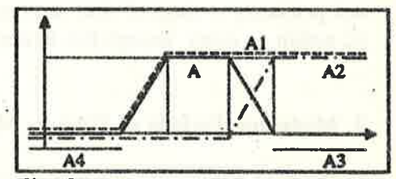


Fig. 3

A3 is the crisp set of time points that *certainly* come after A, while A4 is the crisp set of time points that *certainly* do *not* come after A (Fig.3).
Examples can be found in anterior and posterior relations like *after lunch, before dawn* and *early this morning*.

3.3. Starting point and duration

A third possible model for representing vague time intervals is made up of one fuzzy set for the starting point S of the time interval and one (or more) fuzzy set(s) that indicate(s) the duration D of the time interval (Fig.4). In order to determine the end point E of the vague time interval, several subcases are distinguished below.

The Research of Computer Paradigmatics of the Verb

Natalya Darchuk,
O.O. Potebnya Institute of Linguistics
of the Academy of Sciences of Ukraine
Hrushevsky St., Kiev 1, 252001, Ukraine
Phone: 2282680

Project note

AREA: Automatic morphological analysis and linguistic typology

Summary:
The test of the scientific hypothesis of utilization of the formalized inflective characteristics of the verb for typological analysis of Russian and Ukrainian, and of getting the computer results of the paradigm synthesis.

This work is an attempt of computer ap-proach to the studies of inflexion of verbs at the graphemic level, wihich is stipulated by the purpose of automatic processing of textual information. This approach permits to determine the paradigmatic classes of verbs based on their graphemic structure in the Russian and the Ukrainian languages; to characterize verbal paradigms; to determine the typologic characte-ristics of closely related languages; to create the efficient system for the automatic synthesis of verbal forms.

Исследования машинной парадигматики глагола

Дарчук Н.П.

Резюме:
Проводится проверка научной гипотезы использования формализованных характеристик словоизменения глагола в целях типологического анализа русского и украинского языков. Проводится также анализ результатлв автоматического синтеза словоизменительных парадигм.

Modelling vague lexical time expressions by means of fuzzy set theory

Filip DEVOS [(1)], Nancy VAN GYSEGHEM [(2)]**, Ria VANDENBERGHE [(2)], Rita DE CALUWE [(2)]*

[(1)] Department of Dutch Linguistics
University of Gent (Belgium)
Blandijnberg 2, B-9000 Gent
tel.: +32/9/264.40.82
fax: +32/9/264.41.95

[(2)] Computer Science Laboratory
University of Gent (Belgium)
Technologiepark-Zwijnaarde 9, B-9052 Zwijnaarde
tel.: +32/9/264.55.08
fax: +32/9/264.58.42

Summary (topical paper): On the basis of some large scale inquiries into the nature of time expressions in natural language, in which informants were asked to indicate sharp (crisp) and closed time intervals for a range of linguistic expressions, several models of representing vague lexical time intervals by means of fuzzy set theory are discussed. These models primarily depend on the type of expressions used, and thus take into account the complex heterogeneous semantics of time indications.

Topic area: modelling lexical time expressions - possibility theory - fuzzy set theory

1. Introduction

This paper reports on some large scale inquiries on vague time expressions in natural language. "Time" is a linguistic as well as a non-linguistic notion. Linguistic time is an extremely complex notion, as in natural language different time conceptions and divisions are reflected:
(1) physical or natural time as a fact of extra-linguistic reality. Astronomic notions are often reflected in lexical items (e.g. *year, day, night, noon, season*) which structure and categorize physical time.
(2) artificial or calendar time as the time we can measure and (conventionally) express in lexical items, and we can structure by means of a finite (duo)decimal numerical system (e.g. *hour, century, quarter, minute, week*). In theory, time can be rendered very precisely by means of unique proper names or numbers (e.g. *on Thursday December 24th 1994 at 23h 59min and 59sec*).
(3) experiential or psychological time as the time we experience. Our time conception is not only determined by divisions based on natural phenomena, artificial corrections of these phenomena, and artificial divisions themselves, but also on human experience with time. These "experiential" facts are either culturally or individually determined (e.g. 'week' in our tradition of the five-day week). These three conceptual levels are all reflected in linguistic time, and though they correlate to some extent, it is preferable to clearly keep them apart.

2. Time expressions and vagueness in degree

Lexical time expressions have a complex and heterogeneous semantics, ordinarily showing some degree of lexical vagueness. Representing these natural language expressions for building conceptual models, for integration in database systems and other AI-applications is thus made difficult by a multitude of factors. In this article an analysis is given of a formal means of representing vague lexical time expressions by means of fuzzy set theory, probability theory and

---

* Scientific Associate of the Fund for Joint Basic Research (Belgium)

** Research Assistant of the National Fund for Scientific Research (Belgium)

3.3.1. S is crisp and D is crisp
In this case D should be added to S to determine the (crisp) end point E. Reversely, the duration D can be found by subtracting S from E. For instance, we know that a football game has a duration of *1,5 hours* and that it started *at 8 p.m.*, or we know that a *20 minute* presentation at a conference starts *at 4 p.m.*

3.3.2. S is crisp and D is vague
Here, the end point of time E is represented by means of a fuzzy set, obtained by adding D to S (fuzzy addition through Zadeh's extension principle (Dubois and Prade, 1980)). The membership function for E then equals the membership function for D, shifted over a 'distance' S along the time axis (Fig.5). The duration D can be determined again via the difference E - S.
Examples are: "It rained for *several weeks* after *August 1st*", "*several years* after *the first world war*", or "the *10 o'clock* ceremony lasted *about 2 hours*".

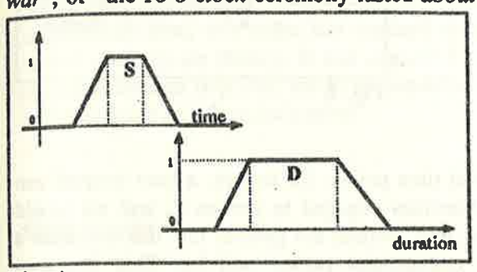


Fig. 4

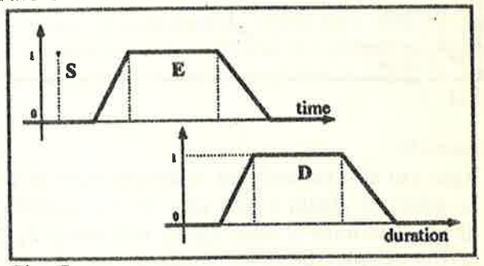


Fig. 5

3.3.3. S is vague and D is crisp
Here, the end point E is also represented as a fuzzy set obtained by adding D to S (fuzzy addition). The membership function for E then equals the membership function for S shifted over a 'distance' D along the time axis (Fig.6). However, in this case, the duration D cannot be determined again from S and E (E is determined as S + D, but E - S does not always reproduce D again; in this case, E - S yields a fuzzy set).

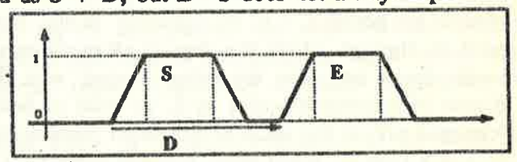


Fig. 6

For instance, a *20 minute* conference presentation can be scheduled *at the end of one or other session*, a *football game* can start *after 8 p.m.*, a new president can be elected *during the first month after someone's dismissal*, a new world record of *10.8 seconds* can be run *in the afternoon*, or a city may have been bombed *for 10 days last month*.

3.3.4. S is vague and D is vague
This model splits up into two submodels:

A. D does not depend on the starting point of time.
The end point E too can then be obtained via fuzzy addition of D to S (for every possible starting point a corresponding vague end point is determined, and all obtained vague end points are reduced via Zadeh's extension principle to one vague end point of time only):

$$\mu_E(t) = \sup_{\substack{x,d \\ t = x+d}} \min (\mu_S(x), \mu_D(d))$$

For example, we know the tennis finals started on *Sunday evening* and took *almost 4 hours*.
In this case, however, the duration D cannot be recalculated from S and E, as E - S yields a too broad fuzzy set, i.e. a fuzzy set the support of which is larger than the support of D.

B. D varies according to the (starting) points of time.
In this case there exists a dependency between the starting point of time and the possible duration of the time interval. Not one fuzzy set is given for the possible duration of the time interval, but two or more fuzzy sets, possibly even a whole series of fuzzy sets, when the duration continuously varies according to the starting point of time (Fig. 7 and 8). The following example may illustrate this. A given person P_1 has age B (e.g. *around 30 years*). This is represented by a fuzzy set with degree of membership μ_B.

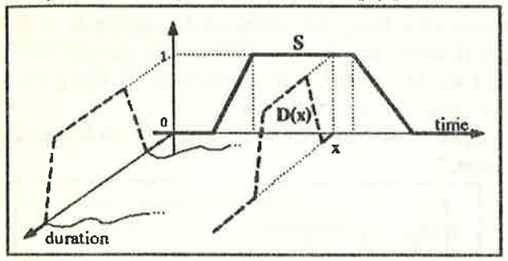


Fig. 7

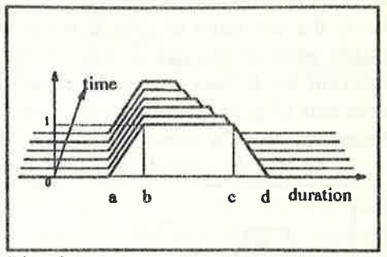


Fig. 8

Remark:
Ages can also be analyzed as special cases of (relative) time points. To an age, a time interval can be added to obtain a new age. If, for instance, P_1 becomes engaged to person P_2 and we would like to determine at what age P_1 will marry P_2, taking into account the general fact that the older a person is who becomes engaged, the shorter the engagement period will be. This fact thus expresses a dependency of the duration of the time interval on the age of the person. This dependency can be formally expressed by a two-dimensional function $\mu_D(x,d)$, with the parameters age X and duration D. By applying Zadeh's extension principle a fuzzy set E can be determined, that describes the age of P_1 at the time of marriage: $\mu_E(t) = \sup_{x,d,\ t = x+d} \min (\mu_S(x), \mu_D(x,d))$

This, however, no longer can be seen as a fuzzy addition. Other formulas for obtaining the end point of time are possible, e.g. by replacing 'minus' by a product, through which the degrees of membership in the fuzzy set $\mu_D(x)$ are being 'scaled' with the degree of membership of x in S, in stead of being 'chopped off' at the value of degree of membership of x in S (Fig.9). When every possible duration D is represented by means of a trapezoidal fuzzy set, the two-dimensional membership function μ_D can be replaced by 4 one-dimensional functions, that, for instance, indicate the course of the 4 breaking points $a-\alpha, a, b, b+\beta$ of the trapezoidal fuzzy set in function of the age x.

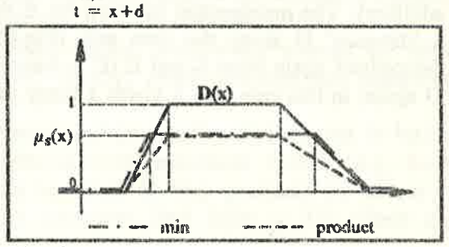


Fig. 9

In this case, it is certainly not possible to recalculate the duration D of the time interval from points S and E.
Imagine the following situation in which person P_1 says to P_2: "Hurry up! The later we will arrive at the party, the less long we will be able to stay". P_1 and P_2 arrive *somewhere between 9 and 10 p.m.*, and they can stay *for about two up to three hours*. Or, the later a plane takes off, the more time people have for saying goodbye.

3.4. Vague sets of time intervals

This is probably the most general way of representing time intervals. Here, a fuzzy set of intervals is analyzed: $\{ \alpha_1/[x_1,y_1], \alpha_2/[x_2,y_2], \ldots, \alpha_n/[x_n,y_n], \ldots \}$
with $\alpha_i \in [0,1]$, point of time $x_i <=$ point of time y_i, all x_i (and y_i) not necessarily different. In principle all cases so far mentioned can be represented in this form. However, since time constitutes a continuum, the representation by means of a set is not always very appropriate to use.

For instance, person P_1 and P_2 plan a *1 hour* meeting and P_1 has but some *free hours* during the day: free hours = {0.5/[10h,10h15], 1/[12h,12h30], 1/[17h,19h], 0.8/[19h,22h] }; => period of the meeting = {1/[17h,18h],..., 1/[18h,19h],..., 0.8/[19h,20h],..., 0.8/[21h,22h]}).

4. Some final remarks and summary

In this paper the granularity of time expressions is not taken into account. Granularity refers to the time levels people use; it constitutes a rather precise hierarchical system of subordinate and superordinate categories in which different shifts may occur; e.g. *hour → day → week → year → decennium → century*...). For every level of granularity the different representations are possible. The representations mentioned in 3.3.4.B. and 3.4. are of little importance for representing vague lexical time expressions.
In summary then, this paper has outlined different models of representing vague time intervals by means of fuzzy set theory. It was argued that this differentiation is needed if the (combined) data obtained through inquiries are to be modelled into a single fuzzy time interval that is suited as the representation of a linguistic term.

Bibliography

Cleeren, R., R. Vandenberghe, N. Van Gyseghem, R. De Caluwe (1993), "The Modelling of Vague Predicates Used in Linguistic Expressions by Means of Fuzzy Set Theory", in: *Proceedings of the Fifth IFSA World Congress, Seoul, Korea, July 4-9, Volume I*, pp. 54-57.

Devos, F. (1993), "Semantische vaagheid en de traditionele woordklassen"["Semantic vagueness and the traditional word classes"], in: *Handelingen van de Koninklijke Zuidnederlandse Maatschappij voor Taal- en Letterkunde en Geschiedenis, XLVI*, pp. 30-45.

Devos, F. (1993b), "Linguïstische problemen bij de incorporatie van vage lexicale tijdsuitdrukkingen in databanken" ["Linguistic problems in incorporating vague lexical time expressions into databases"]. Internal report. University of Gent: Department of Dutch Linguistics.

Devos, F. (1994), "Still Fuzzy After All These Years. A linguistic evaluation of the fuzzy set approach to semantic vagueness" (in preparation).

Dubois, D. and H. Prade (1980), *Fuzzy Sets and Systems: Theory and Applications*, New York: Academic Press.

Dubois, D. and H. Prade (1989), "Processing Fuzzy Temporal Knowledge", in: *IEEE Transactions on Systems, Man, and Cybernetics, 19-4*, pp. 729-744.

Van Gyseghem, N., R. Vandenberghe and R. De Caluwe (1993), "Hoe worden vage tijdsintervallen het best voorgesteld?" ["How are fuzzy time intervals to be represented?"]. Internal report. University of Gent: Computer Science Laboratory.

Moscow Students' Word Associations

Vladimir A. Dolinsky,
"Lingua" Cooperative
129345, Moscow, Ostashkovskaya, 9-2-98.
Phone: (095) 475-8384.
E-mail: polikarp@logos.msu.su

Topical paper.

AREA: Psycholinguistics, semantics, epistemology.

Summary:

Quantitative analysis of data (sample of 101000 associated words: 1010 subjects, 100 stimuli - nouns and adjectives of Russian), obtained in 1991-1993. Some new dependensies and regularities was found.

Mass word association experiment is conducted by author in order to register, collect and proceed experimental data, construct and investigate database and dictionary (thesaurus) of Moscow students' word associations (MCA). The project is carried out within the framework of research program of "Lingua" Cooperativ in colloboration with the M.V. Lomonosov Moscow State University. Program of experiment was conditioned by tasks of wide interdisciplinary investigation (theoretical basis of quantitative analysis, selection and characteristics of stimuli, requirements to subjects (respondents), conditions of test conducting, methods of procedure and treatment of word association data). All subjects were Moscow students, mother tongue is Russian. Number of ss (N) is 1010. The matherial treated up to now involves first 100 distributions of 200, compiling total database and includes 127 nouns and 73 adjectives of Russian. List of stimuli represents different spheres of lexicon (polysemy, frequency, semantic fields etc.) and it was compiled with taking into account the data which have been obtained formerly by other authors (for comparison interlinguistic, diachronic etc.).

адский	диван	невесомый	солдат
бабочка	длина	нога	старый (2)
бассейн	длинный	овощной	стол
белый	добрый	одинокий	субботник
бессмертие	догма	октябрьский	сыр
библия	доктор	паспортный	теннис
близкий	достоинство	паук	телка
близость	духовка	пионерский	товарищ
бог	жестокий	письмо	тонкий
бумага	жизнь	право	тонкость
буря	истина	префикс	углерод
вафля	калитка	примета	хлеб
вкусный	картина	пристальный	цветок
владение	клюква	радость	цирк
влюбленный	кожа	район	цветок
внештатный	красно-речивый	религия	чекист
всесоюзный	красный	рок	черепаха
высокий	ленинский	ручной	черный
глубина	магический	русский	чистка
гобой	матерный	свеча	чистота
гражданин	менестрель	скупой	широкий
демократ	мир	слава	мотки
день	митинг	смертельный	шомпол
деревня	наличие	соблазн	шут
джазовый	невезучий	советский	юбка

Word association test data subjected to computer-assisted treatment present lists of associations with indices of frequency ascertained to the each stimulus (of all word forms reduced to lexeme). Example: stimulus RADOST' (joy).

РАДОСТЬ -

76 счастье	15 встреча	6 неожиданная
74 улыбка	14 светлая	6 удача
65 жизни	13 любовь	6 хорошо
55 смех	13 моя	5 восторг
42 горе	10 гадость	5 гордость
34 грусть	10 печаль	5 огромная
20 веселье	9 собачья	5 победы
20 свет	8 великая	5 праздник
18 большая	8 красное	.....
16 солнце	7 бытия	etc.

After separating of quantitative characteristics distribution of responses to particular stimulus is tabulated and then arranged in order of descending frequency and presented in table as following set of numbers (stimulus RADOST'):

i	F_i	m_f	i	F_i	m_f
1	76	1	13-14	13	2
2	74	1	15-16	10	2
3	65	1	17	9	1
4	55	1	18-19	8	2
5	42	1	20	7	1
6	34	1	21-23	6	3
7-8	20	2	24-28	5	5
9	18	1	29-34	4	6
10	16	1	35-53	3	19
11	15	1	54-104	2	51
12	14	1	105-295	1	91

There was received some interesting results elusidating undecided and vexed questions of lexical associativity in quantitative aspects and the role of word association in language. As well as were found some dependencies and regularities disguised before.

1. The inferential formula (1) which expresses the inversely proportional relationship of ranks and frequencies of responses to particular stimulus appears to be much more complex when the parameter, "b", of distribution curvature is taken into account

$$F_i = \frac{S}{Q}\, i^{-(a+b\ln i)} \qquad (1)$$

(where $F_1 = \frac{S}{Q}$; S - size of sample).

The fitting of formula (1) showed that empirical distbribution followed it. The wide variation of primary response frequency, F_1, and the moderate variation of response assortment (A-vocabulary), L, affects the homogeneity (curvature) of distributions.

2. There is ascertained the dependency of rank-frequency (F_i, i), frequency-spectrum ($\bar{m}_f$, $\bar{F}$), vocabulary-spectrum ($\bar{L}_f$, F) distributions (mean for array).

$$\bar{F}_i = \frac{S}{Q}\, i^{-1}; \quad \bar{L}_f = \frac{S}{Q}\, F^{-1}; \quad \bar{m}_f = \frac{S}{Q}\, [F(F+1)]^{-1} \qquad (2)$$

Thus we may with the help of general formula (2) represent various aspects of quantitatuve structure of mean associative field by natural series of numbers (n), that is common for frequencies (F), sizes of equifrequencial groups (m) and vocabulary (L) of units with frequency not lower than n, and also parameter Q, connected with the size of sample S, i.e. frequency F of association ranked n is equal to the size of vocabulary L of associations with frequency not lower than n and is equal to the quantity m of units with frequency n multiplied by n+1.

$$\bar{F}_n = \bar{L}_n = \bar{m}_n(n+1) \qquad (3)$$

Analysis of concequences of formula (3) may be the subject of examination in quantitative linguistics. Let us remark that from this formula followed, for instance, that the "mean" noun elicits such an assortment of associations, which is quantity of a group united with the primary response woken up by it.

Parameters of mean distribution for array are: $\bar{F}_1$ = 145,06 [35...617]; $\bar{L}$ = 273,05 [158...420]; $\bar{m}_4$ = 191,02 [106...293]; mean number of zero reaction: "-" = 94,14 [26...367]. Parameter Q (for array) = 6,58.

3. Some of dependencies in structure of word association fields is regarded: a) integrated distributions; b) regularities of relative frequencies; c) sigma-distr. and R-distr., that allow to monitor the dynamics of parameters in relation to the size of samples, repeated tests, changes of groups etc; d) matrices of frequency spectra, that group the associations according to other criterion than rank.

4. Repeatedly conducted test evokes more deviations in distributions tnan changes of groops of ss.

5. The spectrum-frequency distribution is not become stable (except as a regularly decreasing rate of diversity) upon increasing the size of the sample (S = 31, 63, 126, 252, 505, 1011, 2022).

6. Dependencies were found between associative characteristics of words in quantitative aspects and some set of its sistemic-linguistic characteristics (frequency in common usage, polysemy, part of speech, abstractiveness etc.).

7. Data of tests conducted with children 6-7 years old were compared with the data obtained from adults using the same stimuli.

8. Data of tests conducted with Russian (MCA and other) and American (Minnessota norms, California norms) students were compared over the arrays.

9. Of special interest was the possibility to analyse semantic characteristics of lexical units by the methods of quantitative linguistics without ad hoc logical hypotheses.

The main linguistic unit, a word, is regarded as embodied in its associative field - an obvious image of unclosed, orded in hierarchy, specific multitude of other words connected with given one by associations inherent to subject and cultural-lingual society (group). An empirical analogue of word association field is some group of associates, obtained in test conducted with native speakers according definite method that allows us to model potential distribution of associative field for a word, a language, a socium.

Whatever heterogeneous word associations might be, the same sense they based on, both the one and diverce, revealing to conscience from varrious sides. Association arises as reaction to outward or inward stimulus, as response on a question, as brightening, clewing, recalling or uncovering of veiled, vague, uncertain or ambiguous sense. Words of a language become revealed in two sides: associative-symbolic, when they are keys opening the entrance into uncounscious, and sign-discoursive, when they become an elements the logical calculation built of. A word is not only an empty element of speech syntagm or grammatical scheme filled in with "meaning" conditioned by context. A word thanks to energy of its associative potential is capable to form and direct speech channel in spite of context. A word is not a chameleon, but is an ice-breaker. It is meaningful not only in the system or in a context. Its accomodating itself in context is supposed that it is unseparable of its associative polysemanticity which it posesses itself even if the only (simple) meaning is attached to it by context. Vital unity of language is composed of those images of a word which it finds in sacral, philosophic, poetic speech. First of all it conserns to nouns and adjectives thanks to verballess way of reproducing of inner motion of associations.

Language as reality semantic and psychophysical consists of field of associations uniting lingual and extralingual presentments characteristic to its bearers. These presentments connect elements of experience and elements of language in unities, maintained by means of associations and comprehending as senses. Verbal associative links are potentially infinite, opened, hierarchically regulated. Their structure can be presented as dynamic, partially overlapping individual associative fields. The associative field of a word serves as a key to revealing its sense. Understanding the meaning, evaluating the sense of a word always involves consideration of the associations it engenders. To understand a word means to set a weight function, quantitative-discrete "key" on the associative field continuum, the weight function assigning different weights to various sections of the field, or ranging different associations according to their force, stability, value or properties. Recurring connections between words are reproduced in cognitive and communicative processes, senses associating, thereby fixing themselves in language and culture according to structure based on the definite numerical code. Associations between words are transforming into word associations, forming images of word or word distribution. In the mind of a linguistic community - at the logically structurized level - these associations become fixed and eventually transformed into meanings. Meanings are associations fixed in the process of

large number of proper syndromes lead to very a bulky data array;
- a high degree of variety in symptoms and syndromes description and a terminological confusion;
- the absence of sufficient statistics for most items.

We think that all these difficulties may only be overcome by using computer methods of diagnostics, these diagnostics based not only on Data Base systems, but on artificial intelligence and decision support methods as well.

We made the first step in this direction by developing the "Diagnostic Point" program (the description is appended). Its diagnostic algorithm is based on the available statistical data and on the idea of diagnostic significance of the symptoms of inborn-hereditary syndromes. This significance reflects the expectations of the specialist working with the system.

"Diagnostic Point" is the result of collaboration between the Syndromology and Clinical Genetics Group (S.S.Rudakov, Doctor of Medicine) of the Russian Medical University and the Systems Analysis Institute, the Russian Academy of Sciences (Yu.A.Dubov, Doctor of Sciences).

We believe that the system has at least two useful properties:
- judging from the results of clinical tests, its diagnostic abilities are very good;
- it incorporates original mathematical ideas and it is not an expert system in the classical sense of the word.

We envisage practical implementation of the system for the following purposes:
- diagnostics of hereditary and congenital syndromes;
- consultations in medico-genetics;
- training medical students in the field;
- research purposes.

4. Notwithstanding the obvious practical benefit of this programme we recognize its limitations. We therefore suggest the following project, which involves development of diagnostic programs for syndromology of a human being, based on the results of linguistic association tests. This project is carried out in collaboration with the "Lingua" Cooperative (V.A.Dolinsky, Doctor of Philosophy).

The need for this approach follows from the fact that the above mentioned reasons for the drawbacks of syndromal diagnostics may be formulated as semantically indistinct description of syndromes. This is due to insufficient knowledge of semantic relations between these descriptions and features of the syndromes and also to the lack of understanding of their linguistic structure.

5. We expect that indistinct description in syndromology possibility may be eliminated. As a result the list of inborn-hereditary syndromes and their indicators may be verified by applying the controlled associative experiment. In this experiment, experts in syndromology and other fields of clinical medicine will have to give verbal responses to word-stimuli related to the informative signs of various syndromes. The rank-frequency and the spectrum-frequency distribution of these associations will reflect qualitative and quantitative parameters of "semantic fields" of syndromes.

We plan to conduct a controlled associative experiment in order to register discrete and/or continuous response to the offered stimuli and present the obtained data with the help of special software imitation of collective associative memory. It is also expected that the associative experiment, will bring us the solution to one of the key problem of artificial intelligence: the integration of the information obtained from different experts.

The methodology and the mathematical apparatus of the association data processing have been elaborated in detail. The wave structure of verbal associations identified recently (V.A. Dolinsky) is of particular interest for the development of algorithms.

The data of the linguistic associative experiment will be used for the construction of the "Associative Thesaurus of Syndromes" database modelled on semantic-associative fields. The system "Associative Thesaurus of Syndromes" will make it possible to carry out differentiated diagnoses of human inborn-hereditary syndromes.

The data based on the above-mentioned algorithm will provide the user with information on both syndromes and symptoms.

Ассоциативный тезаурус синдромов
(Ассоциативный эксперемент в прикладном исследовании)

Долинский В.А.,
Рудаков С.С.

Резюме:
Проект развития диагностической компьютерной системы "Diagnostic Point" для клинической синдромологии, использующий квантитативно-лингвистический подход и ассоциативный эксперимент с целью совершенствования диагностических функций и верификации терминосистемы подъязыка синдромологии.

Detection of spelling errors in Swedish not using a word list en clair

Rickard Domeij* Joachim Hollman* Viggo Kann*
Numerical Analysis and Computing Science
Royal Institute of Technology
S-100 44 STOCKHOLM
SWEDEN

Topical paper

Abstract

We investigate how to construct an efficient method for spelling error detection and correction under the prerequisite of using a word list that is encoded and not possible to decode. Our method is probabilistic and the word list is stored as a Bloom filter. In particular we study how to handle compound words and inflections in Swedish.

Keywords: spelling error detection, spelling error correction, Bloom filter

1 Introduction

How to automatically detect and correct spelling errors is an old problem. Nowadays, most word processors include some sort of spelling error detection. The traditional way of detecting spelling errors is to use a word list, usually also containing some grammatical information, and to look up every word in the text in the word list [6].

The main problem with this solution is that if the word list is not large enough the algorithm will report several correct words as misspelled, because they are not included in the word list. For most natural languages the size of word list needed is too large to fit in the working memory of an ordinary computer. In Swedish this is a big problem, because infinitely many new words can be constructed as compound words.

There is a way to reduce the size of the stored word list by using *Bloom filters* [1]. Then the word list is stored as an array of bits (zeroes and ones), and only two operations are allowed: checking if a specific word is in the word list and adding a new word to the word list. Both operations are extremely fast and the size of the stored data is greatly reduced.

There are two drawbacks to Bloom filters: there is a tiny probability that a word not in the word list is considered to be in the word list, and we cannot store any other information than the words themselves, for example grammatical information.

The word list is stored encoded in a form that is impossible to decode—this is often a prerequisite for commercial distribution. A program that detects exactly the words that are not in the word list can never protect its word list, no matter how it is encoded. This is because it is possible for a modern computer to test, in a few hours, all reasonable combinations of letters and in that way reconstruct the complete word list. This is a crucial advantage of probabilistic spelling error detection methods.

Under the prerequisite of using Bloom filters we have developed a method for finding and correcting misspellings in Swedish texts. The method also works for other languages, similar to Swedish.

*Electronic mail: domeij@nada.kth.se, joachim@matematik.su.se, viggo@nada.kth.se

In this paper we describe the concept of Bloom filters and how it is possible, in spite of the restrictions of the Bloom filters, to handle inflections, compound words and spelling error correction. We discuss the differences between correcting touch-typed texts and optically scanned texts.

2 Swedish word formation

Swedish is a morphologically rich language compared to English. An ordinary verb in Swedish has more than ten different inflectional forms. This makes word listing a heavy task for ordinary computers.

Most words can also be compounded to form a completely new word. For example, the verb *rulla* (roll) can combine with *skridsko* (skate) to form the word *rullskridsko* (roller skate). Since words can combine without limit, it is not even possible to list them. This is a considerable problem for Swedish spell checkers. A great deal of the tiring false alarms that make Swedish spell checkers impractical are compound words.

As the example of Swedish compounding above shows, it is not always possible just to put two words together to form a compound. Stem alteration is often the case, which can mean that the last letter of the initial word stem is deleted or changed, depending (roughly) on what part of speech and inflectional group it belongs to. Between different compound parts an extra *-s-* is often added. However, individual words tend to behave idiosyncratically, thus making compounding hard to describe by general rules.

3 Bloom filters

For a long time, the predominant search method has been *hashing*. The basic idea is to assign an integer to every search key. These integers are then used as indexes into a table that holds all the keys. Ideally, there would be a one-to-one correspondence between the integer indexes and the keys, but this is not necessary and is in fact not even desirable in our application. To achieve good results, it is essential that the function which maps search keys to integers can be quickly computed and that the integers are distributed evenly over all possible table indexes.

If the problem at hand is simply a test for membership (e.g., to check if a word belongs to a word list), then Bloom filters [1] can be used. A Bloom filter is a special kind of hash table, where each entry is either '0' or '1', and where we make repeated hashings into a single table (using different hash functions each time).

A word is added to the table by applying each hash function to the word and entering '1's in the corresponding positions (i.e., the integer indices that the hash functions return).

To check if a word belongs to the word list, you apply the same hash functions and check if all the entries are equal to '1'. If not all entries are equal to '1', then the word was not in the word list.

It can happen that a word gets accepted even if it is not in the word list. The reason is that two different words may have the same *signature*, i.e., '1's in the same positions. Fortunately, the probability for such collisions can easily be adjusted to a specific application. All we have to do is to change the size of the table and the number of hash functions.

Let us compute the probability that a word not in the word list will be accepted by the Bloom filter. Suppose that the word list consists of n words, that the size of the hash table is m, and that we use k independent and evenly distributed hash functions. We would like to compute the probability that the values of the k functions all point to entries equal to '1'.

In the hash table n words have been stored, and for each word k entries have been set to '1'. The probability that a specified entry in the table is still '0' after that is

$$\left(1 - \frac{1}{m}\right)^{k \cdot n}.$$

assuming that the $k \cdot n$ table entry settings were independent. The probability that k random entries in the table all are '1' is

$$f(k) = \left[1 - \left(1 - \frac{1}{m}\right)^{k \cdot n}\right]^k$$

The minimum of this function is found when

$$f'(k) = 0 \Rightarrow \left(1 - \frac{1}{m}\right)^{k \cdot n} = \frac{1}{2},$$

which means that the hash table is used optimally when it is half-filled with ones. We get

$$k = -\frac{\ln 2}{n \cdot \ln\left(1 - \frac{1}{m}\right)} \approx \ln 2 \cdot \frac{m}{n} \approx 0,69 \cdot \frac{m}{n}$$

and the error probability is

$$f(k) = 2^{-k}.$$

Example: *If the word list contains* $n = 100\,000$ *words and we choose* $m = 2\,000\,000$ *as the size of the hash table, we should choose*

$$k = \ln 2 \cdot \frac{2\,000\,000}{100\,000} \approx 13.9 \approx 14,$$

i.e., we should use 14 hash functions in the Bloom filter. The probability that a random word is accepted is $f(14) \approx 6 \cdot 10^{-5} = 0.006\%$.

4 Compounding and inflection

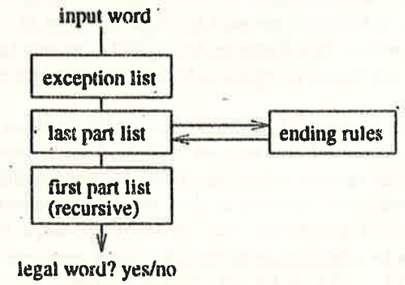
In our program, compounding and inflection are handled by an algorithm that uses a list of ending rules together with three different word lists.

1. the *exception list*, containing words that cannot be part of a compound at all,
2. the *last part list*, containing words that can end a compound or be an independent word,
3. the *first part list*, containing altered word stems that can form the first or middle part of a compound.

Inflection is handled in a straight forward but unconventional way. We are trying a heuristical method to reduce the number of word forms listed, and ensure that all forms of a word is represented. The *last part list* presented above does not actually contain all inflectional word forms. It only contains the basic word forms needed to infer the existence of the rest from ending rules.

Both basic word forms and altered word stems are (semi-) automatically constructed from a machine readable dictionary with inflectional and compound information.

The complete look-up scheme looks like this:



When a word is checked, the algorithm consults the lists in the order illustrated above. In the trivial case, the input word is found directly in the *exception list* or the *last part list*. If the input word is a compound, only its last part is confirmed in the *last part list*. Then the *first part list* is looked up to acknowledge its first part. If the compound has more parts than two, a recursive consultation is performed. The algorithm optionally inserts an extra *-s-* between compound parts, to account for the fact that an extra *-s-* is generally inserted between the second and third compound parts.

The ending rule component is only consulted if an input word cannot be found neither in the *exception list* nor the *last part list*. If the last part of the input word matches a rule-ending, it is considered a legal ending under the condition that the related basic inflectional forms are in the *last part list*. In this way, only three noun forms, out of normally eight, must be stored in the *last part list*. The other noun forms are inferred by ending rules.

Consider the input word *porslinsdockorna* (porslin=porcelain, dockorna=the dolls). The input word cannot be found in the *exception list* nor the *last part list*. Therefore the ending rules are consulted. The following ending rule is found.

-orna ← *-a*, *-an*, *-or*

The rule above is to be read (somewhat simplified) like this: If the words *dock-a* (doll), *dock-an* (the doll) and *dock-or* (dolls) are in the *last part list*, then the word *dock-orna* (the dolls) is a legal word.

Finally the *first part list* is consulted. There the first part of the compound (*porslins-*) is found, thus confirming the legality of the input word.

Our handling of inflections is a possible source of error. For example, the non-existing word *dekorna* can be constructed using the rule above since the words *deka* (degenerate), *dekan* (dean) and *dekor* (scenery) all exist in Swedish. It is important to design the rules in such a way that the number of incorrect words that can be constructed is minimized. There are different ways to obtain better rules. We can include a new suffix on the right hand side of the rule, and at the same time expand the word list with the corresponding inflectional word forms. Another way is to substitute a new suffix for a suffix on the right hand side. A third method is to include a *negated suffix* which works in the following way. If the negated suffix $\mathcal{S}$ is included, and a word exists in the word list with the suffix $\mathcal{S}$, then the rule cannot be applied to that word.

In order to compare different variants of ending rules we generate all possible words that can be constructed from a specific rule. Using the rule in the example above, 1532 words can be generated, and only two of them are incorrect. Thus, the error is $2/1532 \approx 0.0013$.

5 Spelling error correction

Many studies, *see* for example Dameran [3] and Peterson [8], show that four common mistakes cause 80 to 90 percent of all typing errors: transposition of two adjacent letters, one extra letter, one missing letter, and one wrong letter. A method that has proven to be useful for generating spelling correction suggestions is to generate all words that correspond to these four types of mistakes, and see which are correct words.

Words that are generated in this way are said to lie at distance of one from the original word. If there are no correct words within this distance, one could continue the search by increasing the distance by one at each step, but this is of course a very expensive process. An alternative method is described in [4].

This metric is well suited for touch-typed text but other metrics should be used for texts entered in other ways. For instance, hand-written text, and texts that have been entered using OCR-techniques, *see* [9] and the section below, are likely to contain different types of errors.

A problem with the probabilistic method is that when we generate many suggestions for a misspelled word there is a slight possibility that an incorrect word may slip in. It is however possible to reduce such errors to a minimum by introducing a *graphotactical table* as suggested by Mullin and Margoliash [7]. This table holds all allowed *n-grams*, i.e., combinations of n letters,

for some prespecified limit n. We have chosen $n = 4$ and we store the graphotactical table using one bit for every possible 4-gram, '1' if there is a Swedish word that contains the 4-gram and '0' otherwise. A word is accepted as correct only if all its 4-grams appear in the table.

Thus, the reasonableness of the generated words is checked both against the Bloom filter and the graphotactical table. The words that pass both tests will be proposed as corrections.

One should note that the graphotactical table has to be updated if we allow the user to add her own words; fortunately, this is easy.

In earlier studies of automatic spelling correction, *see* for instance Takahashi *et al.* [9], it has been considered impractical to use word lists larger than about 10 000 words. Using our methods, it is possible to have extremely large word lists without sacrificing speed.

6 Correction in optically scanned documents

Correction in connection with OCR is in many ways different from the ordinary spelling correction described above. Not only are we faced with typing errors, but also errors due to imperfections in the text recognition device used. Even a high quality system with a *character* recognition accuracy rate as high as 99% may result in a mere 95% *word* recognition accuracy rate, because one error per 100 characters equates to roughly one error per 20 words, assuming five-character words.

In an optically scanned document we can expect similar looking characters, or groups of characters, such as: 'O'-'0', 'I'-'1'-'l', 'A'-'4', and 'a'-'å'-'ä'-'á'-'à', to cause problems. This is common source of error, especially in a language such as Swedish where 'å', 'ä', and 'ö' are very common "real" letters, i.e., not simply 'a', and 'o' with diacritical marks. Our preliminary results suggest that roughly half of the errors in optically scanned Swedish texts are of this type.

It is natural to choose a metric, i.e., measure of distance between words, different from the one used for (directly) touch-typed texts. In contrast to the usual minimum edit distance, noninteger distances are used here.

Our earlier remarks suggest that this metric depends both on the shape of the characters, and the language (n-gram frequencies). As a step toward fully automatic word correction, or at least in order to help the user of an interactive program, the potential corrections should be ordered by increasing distance from the misspelled word. At the moment, we consider this ranking of candidates to be the most interesting practical problem. The reason for this is that in nearly all cases the correct word is to be found among the candidates, so the real problem is to pick the right candidate. We are currently investigating techniques along the lines of Kernighan et al. [2, 5].

7 Recreating the word list

Any spelling error detection program's word list can be recreated using the following algorithm.

Generate all possible combinations of letters (using the graphotactical table to throw away impossible words) and input them to the spelling error detection program. Note which words the program accepts. These words form the word list.

If the spelling error detection is exact we have recreated the word list exactly, but if it is probabilistic we have got a word list that contains some errors.

If we use the algorithm on our spelling error detection program we will get about 2% nonsense words, which will make the word list useless for others.

This error should not be confused with the probability that a misspelled word is accepted by

the Bloom filter, which is 0.006% in our program.

8 Performance of our method

Here are some notes on the performance of the current implementation of our method. The computer used is a Sun Sparc station ELC, a Unix machine comparable with a fast 486 PC.

- Speed:
  - looking up words in the *exception list* and the *last part list* only: 2500 words/sec,
  - general spelling detection (including compounding and inflection): 700 words/sec,
  - spelling error correction: 20 words/sec.
- Memory requirements:
  - first part list 250 kbyte,
  - last part list 100 kbyte,
  - exception list 25 kbyte,
  - graphotactical table 100 kbyte,
  - ending rules 10 kbyte.

References

[1] B. H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, 1970.

[2] K. W. Church and W. A. Gale. Probability scoring for spelling correction. *Stat. Comput.*, 1:93–103, 1991.

[3] F. J. Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, 1964.

[4] M. W. Du and S. C. Chang. A model and a fast algorithm for multiple errors spelling correction. *Acta Informatica*, 29:281–302, 1992.

[5] N. D. Kernighan, K. W. Church, and W. A. Gale. A spelling correction program based on a noisy channel model. In Hans Karlgren, editor, *COLING-90, The 13th International Conference on Computational Linguistics, Helsinki, Finland*, volume 2, pages 205–210, 1990.

[6] K. Kukich. Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24(4):377–439, 1992.

[7] J. K. Mullin and D. J. Margoliash. A tale of three spelling checkers. *Software–Practice and Experience*, 20(6):625–630, 1990.

[8] J. L. Peterson. A note on undetected typing errors. *Communications of the ACM*, 29(7):633–637, 1986.

[9] H. Takahashi, N. Itoh, T. Amano, and A. Yamashita. A spelling correction method and its application to an OCR system. *Pattern Recognition*, 23(3/4):363–377, 1990.

Word Frequency Distribution in Japanese Text

by

Koichi Ejiri

Ricoh Company Ltd

3-2-3 Shin-Yokohama, Kohoku-ku, Yokohama, 222 Japan

ejiri@ai.rdc.ricoh.co.jp

Abstract

In our last paper [1], we proposed a new parameter, G (=log(N/L){log(N)-1}) , where N is the number of words and L the number of different words, which represents constraint of the target text represented by ASCII code. We found that the same measure is applicable to Japanese text which has no clear word segmentation. By statistical analysis of *kanji*, *kata-kana* and *alphabetic strings*, Japanese text was found to have the similar distribution as English text or computer language. We also introduced a "joint entropy " of *string[i]* and *string[j]* where the latter *string* follows the former *string* after fixed distance. Here, the distance means the number of words(or defined character strings) betwenn *string[i]* and *string[j]*. This entropy is a measure of repetitive description (phrase) in the text.

Introduction

It has been well understood that normal Natural Language Processing technology helps few of the real applications. However, much simpler approach by numerical calculations of natural language has been contributing as useful applications. English error correction by n-gram is one of the earlier example[2]. But the same strategy does not work for Japanese text whose words are imbedded in the continuous character strings. Even for a simple application, such as OCR error correction, heavy grammatical parsing is required[3]. This is still far from real-time text processing of large volume of Japanese text.

In Japanese text, *kanji* has been used for more than 1500 years since it had been introduced from ancient China. In 10-th century, *hira-kana* and *kata-kana* were invented for phonetic expressions. These two *kanas* were used mixed together without clear standard for next 900 years or more. It is only 50 years ago that modern writing style was standardized; newly introduced words from foreign countries are expressed by *kata-kana* and other phonetic description is made by *hira-kana* mixed with *kanji* which represents most of nouns or some verbs. Sometimes, alphabet is directly used in the text recently. As a result, *kata-kana*, *kanji* or *alphabet* strings are commonly observable in modern Japanese text. It is important to note that most of non-*hirakana* strings are nouns.

Considering this situation, Nagao, Mizutani and Ikeda applied a simple rule to pick up a possible keyword by extracting non-hirakana character-string in a Japanese text[4]. So the problem is just to extract these different character sets which are easily classified using their JIS-code expression (JIS=Japanese Industrial Standard). By JIS-code, *kanji*, *hira-kana*, *kata-kana* and JIS style *alphabets* are well classified as is shown in figure 1. Probably, Nagao's work is the first high speed Japanese text processing system using only statistical computations.

Statistical Analysis of Japanese Text

According to our earilier work [1], English is roughly classified in the order of constraint by a parameter G which is defined as

$$\log(N/L)\{\log(N) - 1\}$$

where, N is the number of words and L the number of different words in the target text. In the same paper, it was mentioned that phonetically expressed Japanese text seemed to have the same tendancy. Using similar approach as Nagao, et al., we extracted *kanji*, *kata-kana* and *alphabet* character strings and calculated the parameter G. JIS code expression of Japanese characters are easily identified from ASCII by the first bit of 2 Byte code, which is 1. On the other hand, ASCII code starts from 0 bit. If the code is ASCII, the sentence is easily decomposed into each word [1].

To measure another attribute of a text, we introduced "Joint Entropy" among high frequency character-strings (word for English): If both of the *string[i]* and *string[j]* are within top 5% of frequency distribution, and if *string[j]* follows just after *string[i]*, joint frequency $H(i,j)$ is incremented. Then, we can define Joint Entropy S;

$$p[i,j] = H[i,j]/T,$$

$$S = -\sum_{i,j \subset U} p[i,j] \log(p[i,j])$$

where, U is the set of top-5% of the vocabulary, number T the 5% of the vocabulary.

For *kanji* string measurement, we take only first two *kanjis*, because, each *kanji* easily connects to other *kanji* string to make a new compound word. This two-character kanji string strategy avoides too large vocabulary. Fig.2 shows an example of the frequency distribution.

Before applying this discussion to Japanese text, English texts has been plotted on a graph of parameter G and a new parameter S. The result confirms again our earlier work that G is a measure of constraint of the text(Fig. 3). Here, the texts are categorized as C:computer language, E:expert's knowledge, M:manual, T:technical book, N:novels, W:newspaper and magazine (Table.1).

Figure 4 is the result of the analysis applied to 36 different Japanese texts. Both results, English and Japanese are well segregated by G; lower G suggests normal text while larger G suggests high constrained text. From our definition, S is a measure of repeated use of the same phrase. It is well understandable that S has some correlation with G, because a constraint text tends to have stereotyped or flat expressions. The correlation ratio found to be 0.64(Fig.5).

In English, high frequency words are common words like, *that, is, are, a, it* or *the*, whose connectivity is carefully avoided in sofisticated sentence like newspaper editorials.(Fig.3).

Conclusion

The parameter G ($=\log(N/L)\{\log(N)-1\}$) , where, N is the number of strings and L the number of different strings, is a good measure of constraint for Japanese text if selected sets of character strings are concerned. The joint entropy S , which is defined from its high frequency character strings, suggests repeating expression in the text.

References:

[1] K. Ejiri, A. Smith; "Proposal of a New Constraint Measure for Text", Contributions to Aualitative Linguistics, pp.195-211, Kluwer Academic Publishers edited by R. Koehler and B. Rieger (1993)

[2] C. Suen; "n-Gram Statistics for Natural Language Understanding and Text Processing", IEEE Trans. Vol.PAMI-1, No.2, pp.164-172(1979)

[3]A. Konno and Y. Hongo; "Postprocessing Algorithm based on the Probabilistic and Semantic Method for Japanese OCR", Proceedings of 2nd International Conference on Document Analysis and Recognition", pp.646-649 (1993)

[4] M. Nagao, M. Mizutani, H. Ikeda; "Automatic Keyword Extraction from Japanese Text", Japanese Journal of Information Processing, Vol 17, pp.110-117(1976) (Japanese)

text_name	G	Entropy	P_Entropy	comment
Newspaper and Magazine				
nytedit.806	0.18	0.95	3.91	NewYorkTimes' Editorial(NewsPaper)
nytedit.805	0.18	0.74	3.71	NewYorkTimes' Editorial(NewsPaper)
newswk2.txt	0.14	1.08	4.47	Newsweek(NewsPaper)
newswk1.txt	0.14	0.73	4.28	Newsweek(NewsPaper)
gulfwar.res	0.16	0.92	4.43	NewYorkTimes Article(NewsPaper)
examedit.tex	0.18	1.12	3.87	SF Examiner's Editorial(NewsPaper)
examspor.tex	0.20	1.57	3.66	SF Examiner's SportsArticle
baseball.res	0.15	1.01	5.13	Newsweek Article on baseball
Novels				
hemming.way	0.21	2.66	5.82	"Sun Rises Again" by Hemmingway
anne.nov	0.20	4.17	7.47	"Anne" , novels for young adults
shogun1tex	0.19	4.47	7.68	"Shogun" , a novel
rainbow1	0.21	4.23	7.92	"Rainbow", a novel
Techical Book				
ocronoe.kei	0.23	2.41	4.98	Technical report on OCR
wordhist.asc	0.22	5.53	7.52	Technical report on Natural Language
handbk.ai	0.25	3.82	6.06	Technical Book on AI
analys.mat	0.20	2.96	5.35	Mathematical Book on Analysis
Manual				
msreadme.doc	0.27	9.51	7.95	Manual for C language
msreadme.for	0.27	6.14	6.80	Manual for FORTRAN
exhelp.tex	0.33	3.29	4.54	Help Manual for Expert System
prolog.doc	0.20	1.97	5.89	Manual for Prolog
Expert's Knowledge for Diagnostic System				
pbl4.prm	0.25	4.15	5.80	Expert's knowledge to fix fax problems
Computer Language				
actionf.c	0.34	7.86	4.91	C-code for natural language processing
bs_cache1.c	0.41	15.70	6.18	C-code for Cache memory allocation
bs_dofun.c	0.45	12.97	5.18	C-code for function manipulation
bs_getre.c	0.36	11.61	5.92	C-code for font generation
dwhet.for	0.26	5.55	6.15	Fortran code
demoran.for	0.29	1.37	3.34	Fortran code for demo
demo.for	0.28	0.27	1.50	Fortran code for demo

Table 1.

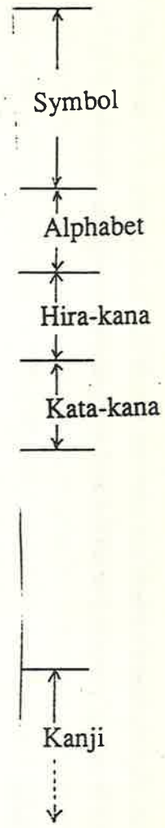


Fig. 1

```
input file name=  murayama_phd1.txt

number of input words  (N =z1)=665
number of different words (L=z2)=316
vocabulary measure        (z2/z1)=0.473684
constraint measure        (z1/z2)=2.111111
   G = log(N/L) / (log(N) - 1) =0.178027
   sum_G, sum_GG, sum_n=11.429858 2.542657 69.000000
   mean_G, sigma_G = 0.165650  0.097717

                freq nor_frq w_length  y  Zipf(k*f**2)

 1        画像   23   34.59     4  0.0730  1196.224  1
 2        離散   20   30.08     4  0.0637   904.517  1
 3          述   15   22.56     2  0.0479   508.791  1
 4        符号   13   19.55     4  0.0417   382.158  1
 5          家   13   19.55     2  0.0418   764.317  2
 6          第   11   16.54     2  0.0355   273.616  1
 7   ディジタル   11   16.54    10  0.0356   547.233  2
 8          2   11   16.54     2  0.0357   820.849  3
 9        多項    8   12.03     4  0.0261   144.723  1
10        幽級    8   12.03     4  0.0261   289.445  2
11        変換    7   10.53     4  0.0230   110.803  1
12        数学    7   10.53     4  0.0230   221.607  2
13        高速    7   10.53     4  0.0231   332.410  3
14        生成    6    9.02     4  0.0199    81.407  1
15   アナログ     6    9.02     8  0.0199   162.813  2
16        分野    6    9.02     4  0.0200   244.220  3
17          量    6    9.02     2  0.0201   325.626  4
18   ベクトル     6    9.02     8  0.0201   407.033  5
19        著者    6    9.02     4  0.0202   488.439  6
20        開発    6    9.02     4  0.0203   569.846  7
21        次元    5    7.52     4  0.0169    56.532  1
22           3    5    7.52     2  0.0170   113.065  2
23        圧縮    5    7.52     4  0.0171   169.597  3
24        文字    5    7.52     4  0.0171   226.129  4
25           1    4    6.02     2  0.0137    36.181  1
26          得    4    6.02     2  0.0138    72.361  2
27           n    4    6.02     2  0.0138   108.542  3
28           m    4    6.02     2  0.0139   144.723  4
29        場合    4    6.02     4  0.0139   180.903  5
30           k    4    6.02     2  0.0140   217.084  6
31        採用    4    6.02     4  0.0140   253.265  7
32        使用    4    6.02     4  0.0141   289.445  8
33        図形    4    6.02     4  0.0141   325.626  9
34   フォント     4    6.02     8  0.0142   361.807 10
35        人類    3    4.51     4  0.0107    20.352  1
36          絵    3    4.51     2  0.0107    40.703  2
37          始    3    4.51     2  0.0108    61.055  3
38        連続    3    4.51     4  0.0108    81.407  4
39        文化    3    4.51     4  0.0108   101.758  5
40          点    3    4.51     2  0.0109   122.110  6
41   ガロワ       3    4.51     6  0.0109   142.461  7
42          体    3    4.51     2  0.0109   162.813  8
43          GF    3    4.51     4  0.0110   183.165  9
44          考    3    4.51     2  0.0110   203.516 10
```

Fig 2

C: computer language
E: Expert's knowledge
M: Manual
T: Technical book
N: Novels
W: Newspaper

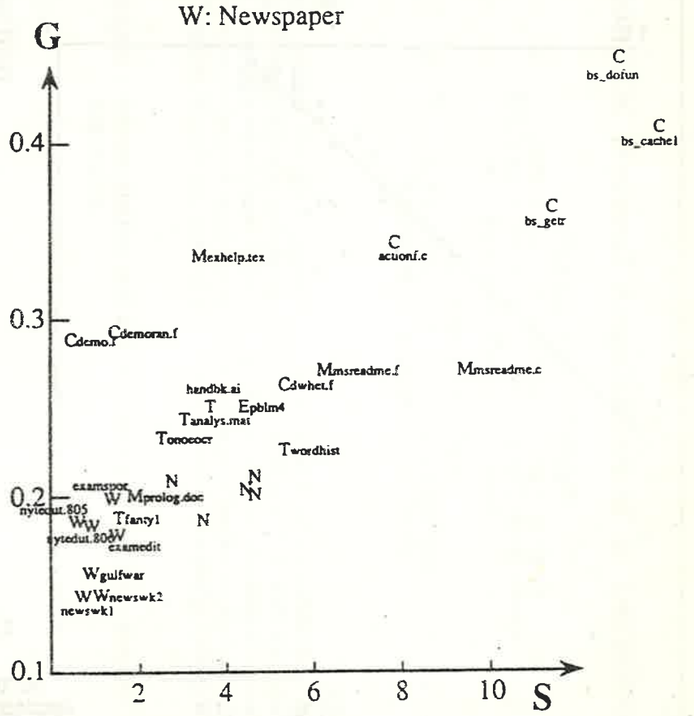


Fig 3

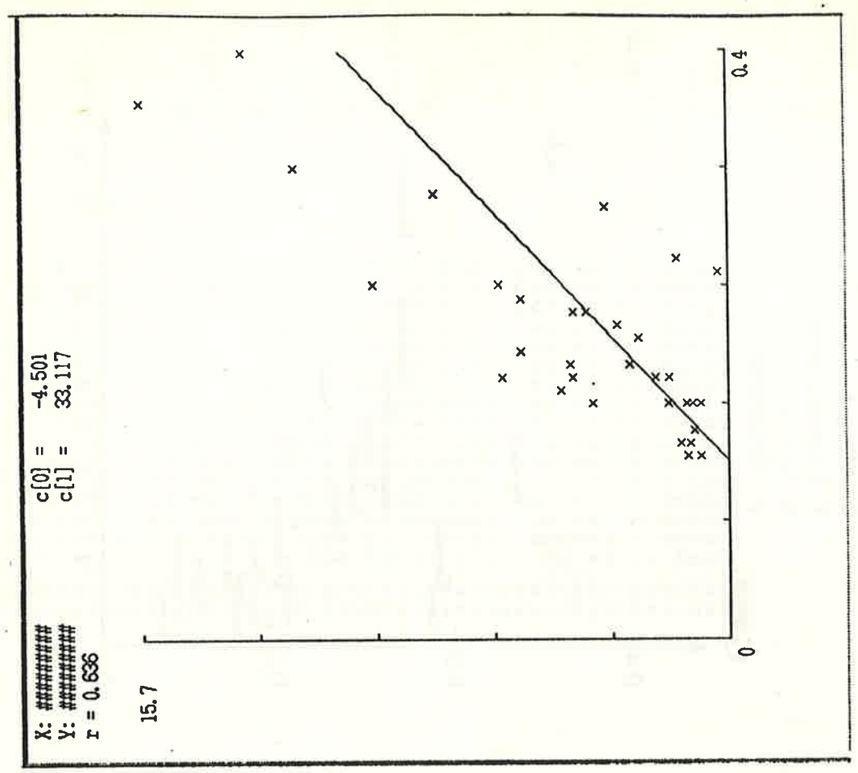
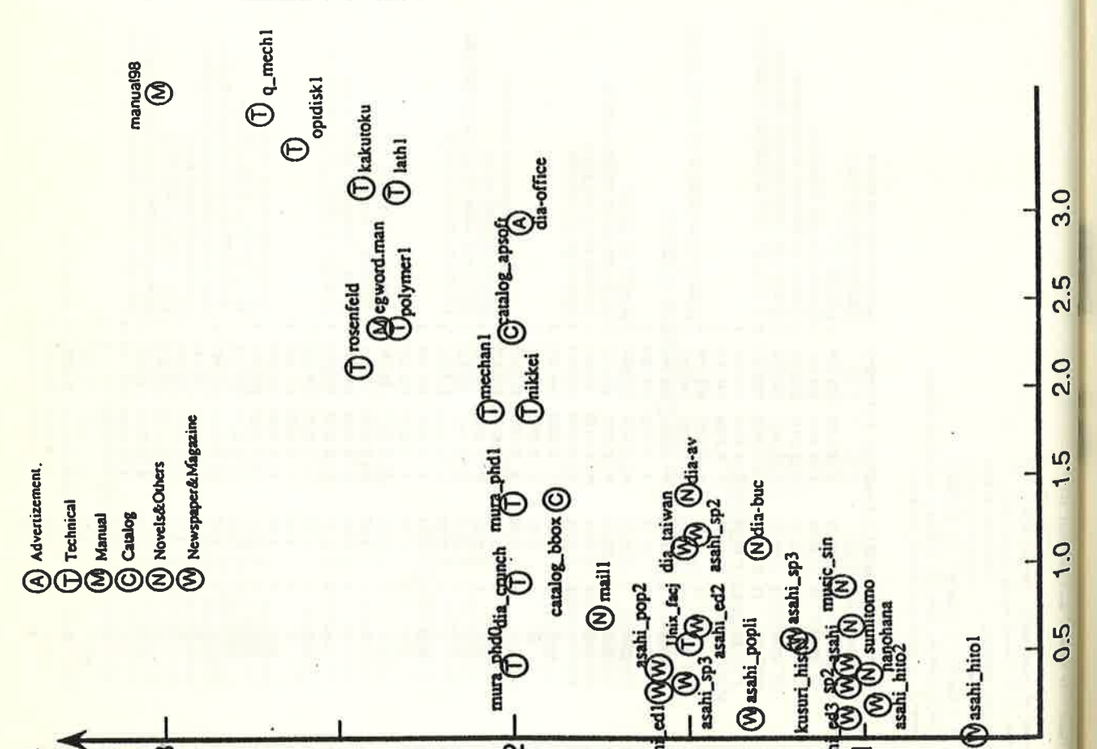


Fig. 5



Automatic Natural Acquisition of a Terminology

Chantal ENGUEHARD
IRIN - Computer Science Research Institute of Nantes
IUT, 3, rue du Maréchal Joffre - 44041 Nantes Cedex 01 - France
enguehard@iut-nantes.univ-nantes.fr

Laurent PANTERA
U.T.C. University of Technology of Compiègne
B.P.649 - 60206 Compiègne Cedex - France

Summary

The authors present the ANA system which is capable of automatically selecting the terminology from a technical domain by the analysis of free text. It uses statistical procedures and a few heuristics which have been inspired by the human learning of a mother tongue. Because the system does not need any syntactical or lexical resources, it is independent of the language (English, French, etc.) and of the level of discourse (technical, colloquial).

Topical paper : documentation and information retrieval, large corpora, knowledge acquisition.

Introduction

The automatic selection of the terminology of a domain has been mainly studied for automatic indexation. BETTS showed that the best quality systems were those which have thesaurus' disposal [BETTS 91]. Also, there is a lack of predefined thesaurus in many technical or scientific new domain, and building a thesaurus requires the participation of specialists of the domain and of terminologists. This a costly work, strongly depending on the willingness of specialists of the domain.
In this paper we present a new approch called ANA 'Automatic Natural Acquisition'. In the first part we shall present our main ideas, a heuristic to modelize the ability to learn a mother tongue, and some tools to recognize the different forms of a unique information. The second part will include the general architecture of the ANA system, the detail of the procedures we created according to our specifications. In the last part we shall show some results and evaluations.

I - Main ideas

Our goal is to define a new way to automatically select the terminology of a domain. These **elements of the terminology**, that we call **terms**, could be used to index the texts or to build a taxonomy, and also in other tasks of natural language processing as desambiguation of words or text generation [SMAD 90].
Note that our convention is to always write terms in capital letters.

I.1 - Specifications

No linguistic knowledge

First, the system should have the ability to treat any text, even if it is not well-written. Actually, there is an industrial need of automatic systems capable of dealing with technical texts, but also with interviews (in a knowledge feedback stage for instance). In these technical texts, **correct syntactical structures are not always adhered to**, and **neologisms frequently occur**[1]. In addition, the huge quantity of texts does not allow for

1 According to [FALZ 89] we say that these texts are written in an operative language. These languages are characterized by the lack of synonymous and the strong use of the enunciative form.

some links between terms are automatically generated, and we intend to transform the set of terms into a semantic network.
The three lists previously determined by the 'Familiarization' module solely constitute the required knowledge to discriminate the terms from the free texts.

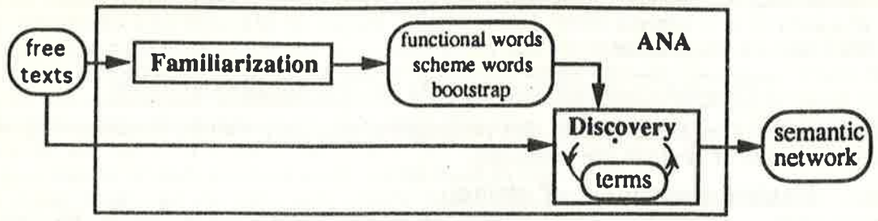


fig. 1 - General architecture

II.2 - Stages of discovery

We now define the three procedures to qualify strings as new terms. We rely heavily upon the postulate which we defined previously (i.e. «frequent co-occurrences of facts are semantically significant.»).

- We consider that **two facts co-occur** when they are separated by less than a fixed number of terms or words. In such a case they are said to be in the same window.
- 'Co-occurrences of facts' will have different **interpretations**. It could be :
  - **two terms** for the term extraction by 'expression',
  - a **term** and a **'scheme word'** for the term extraction by 'candidate',
  - a **term** and a **word** for the term extraction by 'expansion'.

The presentation of these three cases will be illustrated by examples in English on 'Do It Yourself' domain, with :
- W_{fonc} = {"a" "any" "for" "in" "is" "may" "of" "or" "the" "this" "to"},
- "of" is a 'scheme word'
- "WOOD" "COLOUR" "BEECH" "TIMBER", "DIESEL", "ENGINE" are some terms

• 'Expression'

A new term is qualified in the expression manner when two existing terms appear frequently (threshold T_{EXP}) with almost the same arrangement. The most frequent arrangement becomes a new term (and is inserted in a semantic network).

example:
Here are some items of free texts in which the two terms "DIESEL" and "ENGINE" have been identified in the same window :

... "the" ["DIESEL" "ENGINE"] "is" ...

... "this" ["DIESEL" "ENGINE"] "has" ...

... "a" ["DIESEL" "ENGINE"] "never" ...

Result : "DIESEL ENGINE" is qualified as a new term. It is linked to "DIESEL" and "ENGINE" (fig.2).

• 'Candidate'

A new term is qualified in the candidate manner when an existing **term** appears frequently (threshold T_{CAND}) with a **word** that links a **'scheme word'**. This word then becomes a new term.

example:
Here are some items of free texts in which appear in the same window some terms ("WOOD" "COLOUR" "BEECH" "TIMBER"), the word "shape", and the 'scheme word' "of" between them :

... "any" ["shade "of" "WOOD"] "could" ...

... "this" ["shade "of" "COLOUR"] "is" ...

... "the" ["shade "of" "BEECH"] "may" ...

... "new" ["shade "of" "TIMBER"] ...

... "same" ["shade "of" "WOOD"] "in" ...

Result : "SHADE" is qualified a new term.

• 'Expansion'

A new term is qualified in the expansion manner when an existing term appears frequently (threshold T_{EXPA}) with the same succession of words. This succession should not include any term, or any 'scheme word'. The beginning and ending of the new term should not be 'functional words'.

example:
Here are some items of free texts in which appear in the same window the same term "WOODS" with the same term "soft" :

... "use" "any" ["soft" "WOODS"] "to" "make" "this" ...

... "buy" "this" ["soft" "WOODS"] "or" "plastic" "for" ...

... "cheapest" ["soft" "WOODS"] "comes" "from" ...

Result : "SOFT WOODS" is qualified as a new term and inserted in the semantic network with a link to "WOODS" (fig.2).

• Semantic network

In the semantic network we represent some morphological relations, and also the co-occurrence of terms. "DIESEL" and "ENGINE", for example, could occur together but not as "DIESEL ENGINE" : in the sentence *«we need diesel for the engine»* it could be interesting to disambiguate easily the word "engine" by taking in account the proximity of "diesel".

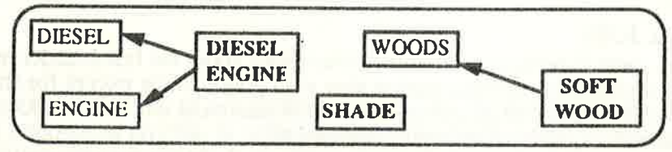


fig. 2 - The semantic network framed by terms

• Incremental process

The discovery module is **incremental**. The system carries on analysing the texts until it does not find any new term. Here are the different stages of treatment for each text :

1 - The system reduces the text by replacing all the signs which are not letter or digit by the blank letter. This 'reduction' step avoids problems which arise due to badly punctuated texts (as ours were), and makes the system both simple and easily maintenable.
2 - A lexical analysis is performed to recognize the terms previously discovered.
3 - The system collects some items of text and memorizes them in the relevant objects ('expressions', 'candidates' and 'expansions').
4 - These items of text are analysed in order to discover new terms and include them in the semantic network.
5 - If new terms appeared the text is analysed again (step 1), if not it stops.

III - Results

English texts

Here are some results obtained by the analysis of scientific papers about acoustics.
Because we had only 25,000 words of text we could not perform the 'familiarization' step and had to give to the system the lists normally provided by this module:

• Data : 2 'scheme words' ("of", "of the"), 34 'functional words', 29 elements in the 'bootstrap' ("ARRAY", "BEAMFORMING", "BOILING", "BUBBLES", "EXPERIMENTS", "IMAGE", "TRANSMISSION", "PULSE", "LEAKS", "TEMPERATURE", "MONITOR", "INSTRUMENTATION", "REACTOR", "RMS", "RECOGNITION", "SCANNING", "SUBASSEMBLIES ", "IMPULSIVE ", "SIGNAL", "TRANSDUCERS", "SOUNDS", "STRUCTURES", "SENSORS", "LOCATION", "SODIUM WATER REACTION", "SGU", "ULTRASONICS", "VELOCITY", "WAVEGUIDE")

• Results: 200 new were found. Here is a sample :

ACOUSTIC ACTIVITY	ACOUSTIC AMPLITUDES
ACOUSTIC BOILING NOISE DETECTION	ACOUSTIC LEAK DETECTION

ACOUSTIC LEAK DETECTION SYSTEM	ACOUSTIC PULSE
ACOUSTIC SIGNAL AMPLITUDE	ACOUSTIC SOURCE LOCATION
ACOUSTIC TRANSMISSION IN SGUS PLANT AND LABORATORY MEASUREMENTS	ACOUSTIC SURVEILLANCE TECHNIQUES FOR SGU LEAK
ACOUSTIC SURVEILLANCE	ADVANCED SIGNAL PROCESSING
AMPLITUDE	ATTENUATION
ATTENUATION IN X CELL	ATTENUATION OF ACOUSTIC SIGNAL
BACKGROUND NOISE IN A SGU	BEEN SET
BEST ESTIMATED SOURCE LOCATION	DIAMETER OF THE SUBASSEMBLY
DISTANCE TRAVELED IN CELL	DROP VELOCITY
DURING A REACTOR SHUTDOWN	ESTIMATE
ESTIMATED SOURCE LOCATION	EXPERIMENT HAS SHOWN
EXPERIMENTAL ARRANGEMENT	EXPERIMENTAL MODEL
EXPERIMENTAL PROGRAMME	FAST REACTOR
NUCLEAR FAST REACTORS	NUCLEAR REACTOR
PATH LENGTH THROUGH A CELL	PATTERN RECOGNITION
SIGNAL AMPLITUDES	SIGNAL ATTENUTION
SIGNAL PROCESSING	SIGNAL PROCESSING TECHNIQUES
SIGNAL STRENGTH	SIGNAL TO NOISE RATIO
VELOCTIY OF SOUND	VELOCITY OF SOUND IN SODIUM

Of course, all are not proper terms : "BEEN SET" or "EXPERIMENT HAS SHOWN", for instance, are bad. However, an evaluation has shown that specialists in this particular domain would keep at least three quaters of these terms.
In addition, such a list of terms can be quickly corrected, as compared to the time which would be involved in the manual selection of terms from free texts.

French texts

We carried out some experiments on French interviews about the fast-breeder reactor Super-Phenix. We performed the familiarisation step with good results except for the 'functional words' list in which we had to add some terms. This treatment will have to be improved. All the other initializations were automatically performed.

- Data : Texts : 120,000 words in texts, 100 'functional words', 6 scheme words', 125 terms in the 'bootstrap'.
- Results : more than 3000 new terms.

These results have previously been published in [ENGU 92] annex 5 and 3,

Conclusion

The ANA system selects the terms about any technical or scientific domain. It is specialized in large corpora which has poor quality because it learns about the language used in texts through an induction process. The texts have not not be corrected, there is no need for any syntactical parser or lexicon,.

Currently we are using the same 'natural' approach to extract some semantic knowledge from networks build by the ANA system (see [ENGU 92] in annex 4).

Bibliography

[BETT 91] Betts, R., Marrable, D., "Free text vs controlled vocabulary, retrieval precision and recall over large databases", Online Inf. 91, Dec., London, p.153-165, 1991.

[BRAI 71] Braine, M.D.S., "The acquisition of language in infant and child", in C. Reed (Ed), "The Learning of Language", N-Y, Appleton, 1971.

[ENGU 92] Enguehard, C., "ANA, Apprentissage Naturel Automatique d'un réseau sémantique", Thèse de Doctorat de l'Université de Technologie de Compiègne, Décembre, 1992.

[SMAD 89] Smadja, F.A., "Lexical co-occurrence : the missing link", Lit. and Linguist. Comput. (UK), vol.4, n°3, p.163-168, 1989.

[SMAD 90] Smadja, F.A., McKeown, K.R., "Automatically extracting and representing collocations for language Generation", 28th Annual Meeting of the Association for Computational Linguistics, NY, USA, p.252-259, 1990.

[WAGN 74] Wagner, R.A., Fischer, M.J., "The string-to-string correction problem", J. of the ACM, vol.21, n°1, p.168-173, Jan., 1974.

A Logical Representation for the Conceptual Coherence of a Sentence

Elisabeth Godbert **Robert Pasero**

Groupe Intelligence Artificielle, CNRS URA 816,
Parc Scientifique de Luminy,
163 Avenue de Luminy, Case 901,
13288 Marseille Cedex 9, France
E-mail : godbert@gia.univ-mrs.fr

Abstract

This paper aims at the definition of a logical representation for the coherence of a sentence with respect to a set of conceptual criteria, in order to prevent the formulation of conceptually incoherent sentences in a natural language interface. First, we define a model for conceptual knowledge corresponding to the most frequent types of incoherence ; this model is derived from the mathematical theory of sets, its uses relations defined on sets and set cardinality. Then, we define the conceptual representation of a sentence in a logical form, from which we can determine whether the sentence is coherent or not.

Submitted to QUALICO 94, Second Inernational Conference on Quantitative Linguistics, Moscow, Russia, 20-24 September 1994.

Topical paper

Topic area : Using the mathematical theory of sets for conceptual analysis of sentences

1. INTRODUCTION

The objective of this work is the definition of a logical representation for the coherence of a sentence with respect to a set of criteria we call conceptual.
The framework of our research is the French project ILLICO, which aims at developing a generator of natural language interfaces in which sentences are composed in a guided mode : at each step of the composition of a sentence, the system has to propose to the user the only words that can lead to a lexically, syntactically and conceptually well-formed sentence ; in order to do this, the words are selected from linguistic and conceptual knowledge about the world of the interfaced application [7].
This paper deals with the well-formedness at the conceptual level : we want to prevent the formulation of linguistically correct but conceptually incoherent sentences, such as *the rabbit speaks, the mothers of Peter, the average of my mark*, etc.
To represent semantics of sentences, we use a (classical) logical language, whose syntax is described in part 2. Then, we have to encircle the nature of conceptual knowledge it is necessary to take into account to prevent incoherences, and to define an appropriate representation of sentence incoherence.

Much research has been done around the problem of determining the truth value of a sentence in a given situation. It has been proved that a two-valued logical system is not sufficient, because incoherent sentences can be considered neither true nor false. To be able to model incoherence, systems generally use three logical values (true, false and incoherent -also called "absurd" or "undefined"-) and several truth-tables have been proposed (*e.g.* [10] [2] [11]).

We propose here a rather different method to model sentence incoherence.

- First, by means of a few examples, we will show the most frequent types of incoherences and the two conceptual criteria which must be respected in order to prevent these incoherences.
- Secondly, we will define a model which permits us to explicitly express, in a logical form, conceptual knowledge corresponding to the criteria. This model is derived from the mathematical theory of sets, it uses relations on sets and set cardinality.
To express the conditions for using the symbols necessary to ensure that the criteria are respected, we define constraints attached to the relational symbols used for the semantic representation of sentences.
- Finally, we will define the conceptual representation of a sentence : to each sentence *S*, we will associate a formula, built from the semantic representation of *S* and from the previously defined constraints. The conceptual representation of *S* will permit us to determine whether it is coherent or not, by using a classical two-valued logic.

2. A LOGICAL REPRESENTATION OF SENTENCES

We consider the world of the interfaced application as a set **E** of individuals inter-connected by relations, that can be described through natural language sentences.
By using singular or plural noun phrases, sentences express connections between objects that can occur either individually or inside sets. So as has been proposed in [9] and [2], we study relations connecting sets of individuals.
We represent the semantics of these sentences by using the language of logic ; we define :

- A set **R** of relational symbols.
Each n-ary relational symbol represents a verb phrase ; it is defined either from a verb or from a noun or from an adjective. A relation in $\mathbf{E}^n$ is associated to it : *talk(x), send(x,y,z), be-human(x), be-John(x), be-father-of(x,y), be-blue(x)*, etc. ; the arguments *x,y,...* represent the subject and complements of the verb phrase represented by the symbol.

- A set **V** of variables.
Each variable *x* represents a set of individuals, and we note $|x|$ its cardinality.
According to the noun phrase denoting *x*, $|x|$ is more or less well known (*a rabbit, Peter and Paul, three children, books, several professors*, etc.).

- The set $\mathfrak{I}$ of all intervals of the type [a,b], where *a* is an integer and *b* is either *a*, or an integer greater than *a*, or ∞.
For each set *x*, according to the natural language denotation of *x*, there exists a minimal element *g(x)* of $\mathfrak{I}$ such that $|x| \in g(x)$. For example : for a singular, g(x) = [1,1] ; for an indefinite plural, we only know that $|x|$ is greater than 1, so g(x) = [2,∞].
We will use the notation (x|[a,b]) to mean "a set x such that : $|x| \in [a,b]$ ".

- Let **L** be the logical language defined by :
- the vocabulary : $\mathbf{R} \cup \mathbf{V} \cup \mathfrak{I} \cup \{\neg, \wedge, \vee, \Rightarrow, \forall, \exists, ,, (,), |'\}$
- the formulas of the four following types, where f_1 and f_2 are formulas of **L** :

1) $f = r(x_1,...,x_n)$ with $r \in \mathbf{R}$ and $x_i \in \mathbf{V}$
2) $f = \neg f_1$
3) $f = f_1\ c\ f_2$ with $c \in \{\wedge, \vee, \Rightarrow\}$
4) $f = q(x|[a,b])\ f_1$ with $q \in \{\forall, \exists\}$ and $[a,b] \in \mathfrak{I}$

In the following parts of the paper, we limit our study to natural language sentences whose semantics can be represented in **L**. This type of logical language has been used by numerous systems to represent semantics of simple sentences since Montague's work on semantics [8] [6] [2].

Example : The sentence *Students listen to Peter* has the following semantic representation :
$\exists (x|[2,\infty])\ (\text{be-student}(x) \wedge (\exists (y|[1,1])\ \text{be-Peter}(y) \wedge \text{listen-to}(x,y)))$

The definition of this logical language is not the objective of this paper. So, we have chosen it very simple, for it is sufficient to illustrate our study of incoherence described below.

3. WHICH CONCEPTUAL INCOHERENCES ?

We will consider that a natural language sentence is incoherent if it contains a transgression to lexical presuppositions of the world of discourse, *i.e.* if it is in opposition to "obvious" (or "common-sense") knowledge attached to the words used in the sentence.
For example, we consider that the expressions *the rabbit speaks, the mothers of Peter, the average of my mark* are lexically and syntactically correct but conceptually incoherent for it is obvious that rabbits don't speak, that a person has only one mother, and that we need several marks to compute their average. Moreover, we consider that *the rabbit does not speak, does the rabbit speak ?* also are incoherent sentences ; this interpretation of coherence agrees with numerous studies done in the area of natural language logic : it is an established fact that, if a proposition is incoherent, its negation is also incoherent, and so is the associated interrogative form [11].

To prevent the production of such sentences in a natural language processing system, this "obvious" knowledge must be modelled in an appropriate form. In our system, it will be contained in the conceptual model of the application, by means of a set of constraints that express which sets of atomic formulas are conceptually acceptable ; these constraints then define, for each symbol *r*, the conditions for an n-uplet $(x_1,...,x_n)$ to be in the graph of the relation associated to *r*. They correspond to lexical presuppositions commonly attached to the verb phrase represented by *r*.

We limit our study, here, to the two following criteria, which correspond to the most frequent incoherences we find in sentences :

- The *domain* criteria, which prevents expressions such as *the rabbit speaks, the mother of the mountain, the table eats a pencil* ; it imposes that the argument $(x_1,...,x_n)$ of a symbol *r* is included in a special subset of $\mathbf{E}^n$. For example, the argument *x* of *speak(x)* must be included in the set of humans. The incoherence of the sentence *the rabbit speaks* comes from the fact that the sets of rabbits and humans are disjoint. On the other hand, *the child speaks* is coherent because the set of children is included in the set of humans.

So, to prevent this type of incoherence, it will be necessary to distinguish a number of subsets of **E** and to specify their possible inclusion or disjunction inter-relations.

• The *connectivity* criteria, which prevents incoherences due to a misuse of singular or plural, as *the mothers of Peter, the student is numerous, the average of my mark*; it imposes conditions on the cardinalities of the arguments x_i of a symbol *r*. Examples :
- In the formula *be-mother-of(x,y)*, |x| must be smaller or equal to |y| .
- In the formula *practise(x,p,d)*, (*the person x practises the profession p at the date d*), if we assume that a person can practise only one profession at one time, each pair *(x,d)* can be connected to only one *p*. So, if |x|=k and |d|=h, |p| must be smaller or equal to *(k*h)*.

The definition of domain and connectivity constraints will permit us to take these criteria into account.

4. COHERENCE IN TERMS OF DOMAINS

We first define a set **D** of *domains* of **E** to which the individuals belong : a domain *D* is a subset of **E**, intentionally defined by one or more properties, and often related to a natural species (*human, animal, fruit tree*, etc.). **E** is the greatest domain. The set **D** of all the domains is partially ordered by the inclusion relation defined in $\mathcal{P}(\mathbf{E})$ (the set of all the parts of **E**).

We then define a set **Dec** of *decompositions* of domains :
A *decomposition* of a domain *D* is a set $\{D_1, D_2, ..., D_p\}$ of disjoint domains, each strictly included in *D* ; the decomposition is noted : $D >> (D_1, D_2, ..., D_p)$.
We can define several decompositions of the same domain. The coherence of the set **Dec** is maintained by the help of several rules described in [3].

Examples : Animate >> (Human, Animal) ;
Human >> (Man, Woman, Child) ; Human >> (Teacher, Doctor, Farmer, Trader).

Definitions

• We define an application *dom* from **R** into $\bigcup_{n\geq 1} \mathbf{D}^n$ that, to each n-ary symbol *r*, associates an element of $\mathbf{D}^n$, called the domain of coherence of *r*.
• $dom(r) = (D_1, ..., D_n)$ is the domain constraint of *r*.
We will note $dom(r)_i$ the *i*-th component of *dom(r)*.

Example : For the relational symbols *be-rabbit* and *speak*, we define the domain constraints : *dom(be-rabbit) = Animal ; dom(speak) = Human.*
We go back to the sentence *the rabbit speaks*, represented by $\exists(x/[1,1])$ *(be-rabbit(x) ∧ speak(x))*. The incoherence of the sentence corresponds to the fact that there cannot exist a set *x* included in both domains Animal and Human. This is generalized, below, by the definition of the conceptual representation of a formula of **L** in terms of domains, expressed in a logical form.

Definition
For each formula *f* of the language **L**, the conceptual representation of *f* in terms of domains, noted $R_{dom}(f)$, is the logical expression recursively defined in the following way :

1) if $f = r(x_1,...,x_n)$ $R_{dom}(f)$ is : $(x_1 \subseteq dom(r)_1) \wedge ... \wedge (x_n \subseteq dom(r)_n)$

2) if $f = \neg f_1$ $R_{dom}(f)$ is : $R_{dom}(f_1)$

3) if $f = f_1$ c f_2 $R_{dom}(f)$ is : $R_{dom}(f_1) \wedge R_{dom}(f_2)$

4) if $f = q\ (x/[a,b])\ f_1$ $R_{dom}(f)$ is : $\exists x\ R_{dom}(f_1)$

If a sentence *S* has the formula *f* as semantic representation, *S* is said *coherent in terms of domains* if and only if the formula $R_{dom}(f)$ is true (*i.e.* if and only if the set **Dec** of decompositions of domains shows the relations expressed in $R_{dom}(f)$ to be true).

Example : We go on with the above example.
S is the sentence : *the rabbit speaks* ; *f* is the formula : $\exists(x/[1,1])$ *(be-rabbit(x) ∧ speak(x))*.
The domain constraints are : *dom(be-rabbit) = Animal* , *dom(speak) = Human* ;
$R_{dom}(f)$ is : $\exists x\ (x \subseteq Human) \wedge (x \subseteq Animal)$.
From the decomposition *Animate >> (Human, Animal)*, we know that the domains Human and Animal are disjoint. So, $R_{dom}(f)$ is false, and *S* is incoherent.

5. COHERENCE IN TERMS OF CONNECTIVITY

A connectivity constraint on a symbol *r* specifies dependencies between the cardinalities of the sets x_i occurring in a formula $r(x_1,...,x_n)$. The notion of connectivity introduced here for natural language processing is an adaptation of the notions of cardinality and multi-valued dependency defined in the relational database area [1].

Definitions
• For $n \geq 1$, we call *c-triplet* of type *n* any triplet *(s, k, j)* such that :
s is a strict subset of {1,..., n}, $k \in \{1,..., n\}$, $k \notin s$, $j \in \mathfrak{I}$.

• We note $\mathbf{T}^{(n)}$ the set of the c-triplets of type *n*, and $\mathbf{T} = \bigcup_{n\geq 1} \mathbf{T}^{(n)}$

• We define an application *conn* from **R** into $\mathcal{P}(\mathbf{T})$ (the set of all the parts of **T**) : for any *n*-ary symbol *r*, *conn(r)* is a finite set of c-triplets of type *n*, called the connectivity-table of *r*.
• $conn(r) = \{t_1,..., t_q\}$ is the connectivity constraint of *r*.

The relation $(s,k,j) \in conn(r)$ has the following meaning : in the formula $r(x_1,...,x_n)$, if each argument x_i, with *i* in the set *s*, is a set with cardinality 1, then the cardinality of the *k*-th argument must be in *j*.

Examples : For the the relations *be-mother-of(x,y)* and *practise(x,p,d)* (cf § 3) we define the constraints : $conn(be\text{-}mother\text{-}of) = \{(\{2\}, 1, [1,1])\}$, $conn(practise) = \{(\{1,3\}, 2, [1,1])\}$

We generalize the conditions defined for these two examples by the definition of the following formula $\varphi(t, x_1,...,x_n)$, where *t* is a c-triplet :
If t = (s,k,j), s = $\{i_1, ..., i_h\}$ and j = [m,p], $\varphi(t, x_1,...,x_n)$ is : $m \leq |x_k| \leq p * |x_{i_1}| * ... * |x_{i_h}|$
If t = (s,k,j), s = { } and j = [m,p], $\varphi(t, x_1,...,x_n)$ is : $m \leq |x_k| \leq p$

We can now define the conceptual representation of a formula *f* of **L** in terms of connectivity.

Definition
For each formula *f* of the language **L**, the conceptual representation of *f* in terms of connectivity, noted $R_{conn}(f)$, is the logical expression recursively defined in the following way :

1) if $f = r(x_1,...,x_n)$ and if we note $conn(r) = \{t_1,..., t_q\}$
$R_{conn}(f)$ is : $\varphi(t_1, x_1,...,x_n) \wedge ... \wedge \varphi(t_q, x_1,...,x_n)$

2) if $f = \neg f_1$ $R_{conn}(f)$ is : $R_{conn}(f_1)$

3) if $f = f_1$ c f_2 $R_{conn}(f)$ is : $R_{conn}(f_1) \wedge R_{conn}(f_2)$

4) if $f = q\ (x/[a,b])\ f_1$ $R_{conn}(f)$ is : $\exists x\ ((|x| \in [a,b]) \wedge R_{conn}(f_1))$

If a sentence S has the formula f as semantic representation, S is said *coherent in terms of connectivity* if and only if the formula $R_{conn}(f)$ is true (*i.e.* if and only if the information we have about the cardinalities of the sets x_i occurring in f shows all the relations expressed in $R_{conn}(f)$ to be true).

Example
Let S be the expression *the mothers of Peter* ; its semantic representation is the formula :
$f : \exists (y\,|[1,1]) \ (\ be\text{-}Peter(y) \wedge \exists (x\,|[2,\infty]\,)\ be\text{-}mother\text{-}of(x,y)\ )$;
The connectivity constraint is : $conn(be\text{-}mother\text{-}of) = \{(\{2\}, 1, [1,1])\}$
$R_{conn}(f)$ is : $\exists y\ (\ (|y| \in [1,1]) \wedge \exists x\ (\ (|x| \in [2,\infty]) \wedge (1 \leq |x| \leq 1 * |y|)\ )\ )$
So, $R_{conn}(f)$ is false, and S is incoherent.

6. CONCEPTUAL COHERENCE OF A SENTENCE

A sentence will be said (globally) conceptually coherent if and only if it respects all the conceptual constraints we have defined. In the present state of our research, this means that a sentence is said conceptually coherent if and only if it is coherent in terms of domains and coherent in terms of connectivity.
So, if S is a sentence and f its semantic representation, S is said conceptually coherent if and only if the formulas $R_{dom}(f)$ and $R_{conn}(f)$ are true.

7. CONCLUSION

We have described our study of two conceptual criteria, the domain and connectivity criteria, whose respect permits to prevent the formulation of some types of incoherences in natural language.
More examples of conceptually coherent or incoherent sentences in terms of domains and connectivity can be found in [3] and [4].
Our present research aims at improving the conceptual representation of sentences : of course, there exist other forms of incoherence, and we go on our work in that way ; by modelling other conceptual criteria, we will complete and therefore refine the conceptual representation of sentences, to prevent other types of incoherences.
Our system is implemented in Prolog on Macintosh. An application of the system ILLICO has recently been developped : the system KOMBE, a speech aid system for disabled persons [5].

REFERENCES

[1] Aho A.V., Beeri C., Ullman J.D. : The Theory of Joins in Relational Databases. *ACM Transactions on Database Systems*, Vol. 4, n° 3, September 1979.
[2] Colmerauer A. : Un sous-ensemble intéressant du français. *R.A.I.R.O., Informatique théorique*, vol. 13, 1979.
[3] Godbert E. : Les contraintes de domaine dans un modèle conceptuel associé à une interface en langue naturelle. Note interne, GIA, University of Aix-Marseille, 1991.
[4] Godbert E., Pasero R. : La Connectivité en Langage Naturel : Modélisation de Contraintes sur le Nombre. *Journées Internationales d'Avignon sur le Traitement du Langage Naturel*, Avignon, May 1993.
[5] Guenthner F., Krüger-Thielmann K., Pasero R., Sabatier P. Communications Aids for ALS Patients. *3rd International Conference on Computers for Handicapped Persons*, Vienne, Austria, 1992.
[6] Kamp H. : A Theory of Truth and Semantic Representation, in Groenendijk, Janssen & Stokhof (eds.) *Truth, Interpretation and Information*, GRASS 2, Foris, Dordrecht, 1984.
[7] Milhaud G., Pasero R., Sabatier P. : Partial Synthesis of Sentences by Coroutining Constraints on Different Levels of Well-formedness, *COLING Conference*, Nantes, 1992.
[8] Montague R. : *Formal Philosophy*, R. H. Thomason (ed.), Yale University Press, New Haven, 1974.
[9] Pasero R. : *Représentation du français en logique du premier ordre en vue de dialoguer avec un ordinateur*. Thèse de 3e cycle, GIA, University of Aix-Marseille, 1973.
[10] Rescher N. : *Many-valued logic*. New-York, MacGraw Hill, 1969.
[11] Véronis J. : Un modèle logique de l'erreur dans le dialogue homme-machine en langage naturel. *Revue d'Intelligence Artificielle*, vol. 3, n° 1, 1989.

Computer Model of the Suffix Zone in Ukranian

Tatyana Gryaznuchyna,
O.O.Potebnya Institute of Linguistics
of the Academy of Sciences of Ukraine
Hrushevsky Street 4, Kiev 1, 252001, Ukraine
Phone: 2282680

Lyudmyla Alexeyenko,
Taras Shevchenko University
Shevchenko Av. 14, Kiev 1, 252001, Ukraine

Project note

AREA: Automatic morphemic analysis

SUMMARY:
Carrying out of the automatic morphemic segmentator of the Ukrainian language and of the programm for statistic processing of the acquired results.

The formation principles of the morphemic segmentator are set out by the work. The segmentator is an instrument of the system analysis of syntagmatic relations at the level of morpheme system of words of different grammatical classes. Syntagms of the suffix zone are described in the terms of symmetrical and rhythmic structures which represent the combination schemes for the suffixes of different lengt During the automatic text processing the syntagmat characteristics acquire digital expression.

Компьютерная модель зоны суффиксо в украинском языке

Грязнухина Т.
Алексеенко Л.

РЕЗЮМЕ:
Проведение автоматической морфемн сегментации украинского языка и выполнен программы статистического анализа получени результатов.

The Computer Version of the "Part of Speech" Concept

Lyudmyla Alexeyenko,
Kiev, Taras Shevchenko University
Shevchenko Av. 14, Kiev 1, 252001, Ukraine
Phone: 2691684

Tatyana Gryaznuchyna, .
Kiev, O.O. Potebnya Institute of Linguistics
of the Academy of Sciences of Ukraine
Hrushevsky Street 4, Kiev 1, 252001, Ukraine
Phone: 2282680

Project note.

AREA: Automatic morphological analysis.

Summary:
Problems of computer modelling of Russian and Ukrainian morphology are discussed.

The complex meaning of the part of speech in Slavonic lan-guages is mainly formed in post-root zone of the wordform, that's why it can be expressed in the AOT system by the terms of quasi-inflexions and context procedures. The computer version of the part-of-speech classification in Russian . and Ukrainian is observed in the paper.

Компьютерная версия концепта "Часть речи"

Алексеенко Л.,
Грязнухина Т.

Резюме:
Обсуждаются проблемы моделирования морфологии русского и украинского языков

Lexical constraint grammars

Jean-François HÜE From IRIN
(Computer Science Research Institute of Nantes, FRANCE)
3, rue du Mal Joffre, NANTES, cedex 01, FRANCE
email : hue@iut-nantes.univ-nantes.fr
tel : (33)40306052
fax : (33)40306001

Topic

Language theory for quantitative linguistics.

Abstract

This paper presents the concept of **Lexical Constraint Grammar** and its implementation in a **Smalltalk environment**. This concept is useful for realising **context sensitive grammar** syntactico-semantic parsers for large text corpora. We have built a user-friendly software tool for linguists based on this concept.

1. Introduction.

In the natural word, meanings are expressed by forms used in languages. The syntax for these forms may be highly complex, and very irregular. Fortunately, there is one domain, the Indo-European language family, where that complexity and irregularity are greatly attenuated. Three principles are used to this end: the use of a restricted number of graphical symbols, their linear arrangement within texts, and the existence of relatively small numbers of grammatical rules, which remain relatively stable over long periods of time.

Nevertheless the great variety of things that these languages must describe - objects, actions, temporal aspects - makes the syntactico-semantic analysis of natural language texts highly difficult.

One science in particular, **linguistics**, has developed with the aim of studying all aspects of human language. One very rich branch of this wider science is **language theory** [4] [5] [6] [7] [8], which includes among its goals the implementation of the syntactico-semantic parsers required in the construction of compilers for artificial computer languages. This theory, combined with others, is behind the implementations of natural language understanding systems.

Many syntactico-semantic parsers for natural language texts follow strictly the linear presentation of our languages because their design is directly based on the definition of grammars in language theory. In this text, we present **lexical constraint grammars**, in which certain terminals play a special rôle; we will refer to the set of these terminals as the lexicon. Work in this domain has been carried out on speech recognition [11], and has recently been applied to the recognition of syntactic errors [12].

2. Summary of the ideas behind this work.

- A language with a possibly complex grammar will be associated to a lexical constraint grammar, the sub-grammars of which are simpler (algebraic and deterministic [13]).
- Analysis of the text will not be linear, but based around the positions of certain elements of the lexicon. One possible generalisation, which has been the object of further study, is to consider not simply terminals, but complex syntactic structures .
- These lexical constraint grammars may be defined in terms of sub-grammars which may themselves be Language Constraint Grammars, thus allowing a **recursive depth-first syntactico-semantic analysis.**
- The parsers based on the **lexical constraint grammar** principle facilitate the analysis of texts which include sections which are unanalysable - either because the parser is not powerful enough or because there are errors in the text. Analysis tales place locally, 'around' lexical symbols, thus allowing a **partial analysis.**
- **A non-linear combination of n algebraic grammars** allows us to **analyse context sensitive languages.**

3. Lexical Constraint Grammars.

3.1 Definition.

A lexical Constraint Grammar is a set {L, li, G, E, F}, where

L is a set of symbols called the Lexicon

Li is a set of symbols called the **set of bounding symbols.** When li is empty the LCG is said to be **unbounded**, otherwise it is said to be **bounded.**

G is a grammar the set of terminal symbols of which is equal to L.

E is a set of grammars which may possibly be themselves Lexical Constraint Grammar.

F is a function which maps L onto E*E

$$F: L \rightarrow E^*E$$
$$l \rightarrow (G_g(l), G_d(l))$$

where Gg(l) and Gd(l) are the left and right grammars of l respectively.

The notation GL° will be used for the LCG associated to GL where $GL^\circ = \{L, li, E, G^\circ, F\}$ and G° is the grammar which generates the monoide $L^*=UL^n$ on L.

3.2 Example 1.

Consider the language U = {x | x = $a^n b^n c^n$, n N*} which cannot be defined by an algebraic grammar [7]. We will define an unbounded Lexical Constraint Grammar (l1={}) .

GL1 = {L1, Li1, G1, F1} which generates it exactly:

Li1={}

G1={$V1_N$,$V1_T$,A1,P1} with

$V1_N$={A1,B1}

$V1_T$=L1

P1={A1-->?B1?,B1-->bB1,B1-->nil}

where ? represent a symbol different from b.

E1={$G1_g$(b),$G1_d$(b)}

F1:L1-->E1*E1

b->($G1_g$(b),$G1_d$(b))

with

$G1_g(b)=\{G1_{gN}(b),G1_{gT}(b),A1_g(b),P1_g(b)\}$

where

$G1_{gN}(b)=\{A1_g(b)\}$

$G1g_T(b)=\{a\}$

$P1_g(b)=\{A1_g(b)\text{-->}a\}$

with

$G1_d(b)=\{G1_{dN}(b),G1_{dT}(b),A1_d(b),P1_d(b)\}$

where

$G1_{dN}(b)=\{A1_d(b),GL1°\}$

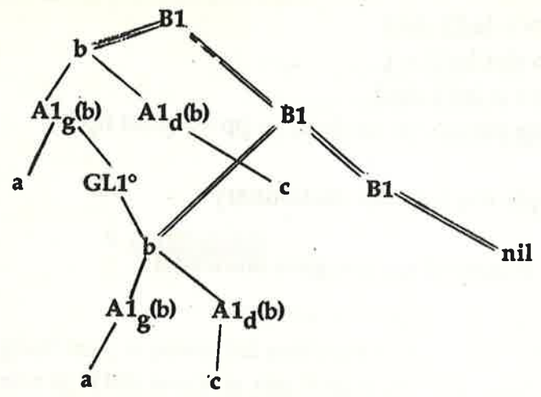
$G1_{dT}(b)=\{c\}$

$P1_d(b)=\{A1_d(b)\text{-->}GL1°c,A1d(b)\text{-->}c\}$

As we can see, the grammars used in this Lexical Constraint Grammar are all algebraic and deterministic and there is a recursive call to the Lexical Constraint Grammar in the left grammar of 'b'.

Thus a language for which there exists no algebraic grammar may be described by a LCG which calls two deterministic algebraic grammars.

3.2.1 Syntactic net associated to 'a a b b c c'.



the dotted line is used to represent the derivations of the left and right grammars of a symbol of the lexicon.

3.2.2 Syntactic analysis of 'a a b b c c'.

The parser begins a left to right search for the first 'b'. A symbol which has been analysed is underlined and is ignored in subsequent steps of the analysis.

a a b b c c search for the first 'b'
a a b b c c leftwards analysis
a a b b c c search for the second 'b'
a a b b c c leftwards analysis

(there are no more b)

a a b b c c rightwards analysis
a a b b c c rightwards analysis

The text is recognised as being grammaticaly correct with respect to GL1.

3.2 Example 2.

We give this example with the syntax of our lexical constraint grammars parser :

Grammaire Lexicale : GLtest
- Portée limitée
- Lex = {vb}
- Lim = {.}

An intelligent Chinese input system using statistical information between words

Sun Da Jiang

sun@nak.math.keio.ac.jp

Junya Tsutsumi, Tomoaki Nitta, Kotaro Ono, Shiho Nobesawa, Masakazu Nakanishi

Nakanishi Laboratory

Faculty of Science and Technology, Keio University

3-14-1 Hiyoshi Kohoku-ku

Yokohama, Kanagawa, 223 JAPAN

Abstract

This paper offers a new method for Chinese input system to translate syllables (phonetic letters, *Pinyin* letters) into Chinese characters without using any grammatical information. Instead, this experimental system uses statistical information between words given in a Chinese corpus and uses a method to input "sentence by sentence". The result shows that this system is intelligent and efficient for most of the sentences.

Topical paper

Topic area: segmenting syllables, quantitative linguistics.

1 Motivation

In Chinese word processor, how to input Chinese characters (*Kanji or Hanzi*) is a great problem to be solved because Kanji is different from alphabet characters.

In order to change Japanese alphabet to Kanji, grammatical analysis between Japanese syllables is adopted. To improve the input system, consecutive syllable analysis has come to be widely used, furthermore semantical methods are used to distinguish the homonym. However, under the present circumstances the technique for Chinese word processor is still in a beginning stage.

This paper describes a method for Chinese input system to translate syllables, which can be represented by phonetic letters (*Pinyin*), into Chinese characters without using any grammatical information. Instead, this system PCS (Pinyin into Chinese characters using statistical information between words) uses the statistical information between syllables to choose better Chinese-like sentences.

With a dictionary, it is not very hard to change Pinyin to Chinese characters, but an intelligent input system should not be "word by word", but should be "sentence by sentence ". The problem is how to decide which sentence written by Chinese characters can make a good match for a sentence written by Pinyin.

This system uses statistical information between words to decide how the unity of syllable string looks like Chinese. So, to make the input system useful, we have to consider how to get the right Chinese sentences from all the possible Chinese-like sentences.

In this paper to get the statistical information between words, mutual information (MI) is used to calculate the relationship between words found in the given sentences. A corpus of Chinese sentences is used to gain the MI. By using this method, we implement system named PCS to experiment on Chinese sentences. The result shows that this system is intelligent and efficient for most of the sentences.

2 What is Chinese standard language ?

Chinese language have some dialects separated from ancient times. Each dialect has its own pronunciation. It is characteristic of Chinese language to express the semantics with the same Chinese character though the pronunciation differs to each other.

2.1 The dialect of Chinese language

The dialect of Chinese can be classified as follows.

- BeiFang (北方) language : on behalf of BeiJing (北京) language.
- JiangNan (江南) language : on behalf of ShangHai (上海) language.
- HuNan (湖南) language : on behalf of ChangSha (長沙) language.
- JiangXi (江西) language : on behalf of NanChang (南昌) language.
- KeJia (客家) language : on behalf of GuangDong province Mei prefecture (広東省梅県).
- MinBei (閩北) language : on behalf of FuZhou (福州) language.
- MinNan (閩南) language : on behalf of XiaMen (廈門) language.
- GuangDong (広東) language : on behalf of GuangZhou (広州) language.

2.2 Chinese standard language

Chinese standard language is a common language of the contemporary Han (漢) Chinese. Standard pronunciation is the pronunciation of BeiJing language. BeiFang language uses typical terms and grammar in representative, famous spoken language. So the basic of the dialect of Chinese language is BeiFang language.

2.3 Pronunciation of Chinese standard language

Pinyin (phonetic letters for Chinese syllables) is a set of Chinese phonetic symbols like Japanese roman alphabets. A syllable consists of phoneme, and the phoneme is classified to consonant and vowel.

The head of phoneme is a consonant which we call ShengMu (声母) in China. After a consonant, there are one or two or three vowels, we call it YunMu (韻母). A vowel consists of at most three letters . The vowel itself can be a syllable.

2.3.1 The structure of consonant

The consonant consist of the following symbols.

- b p m f d t n l g k h j q x zh ch sh r z c s.

2.3.2 The structure of vowel

The vowel can be devided into two parts. One is a single vowel (単韻母) and the other is a double vowel (複韻母).

- single vowel : i u ü a o e er.
- double vowel : ai ei ao ou ia ie ua uo üe iao iou uai uei an en in ün ian uan üan uen ang eng ing ong iang iong uang ueng.

2.3.3 The tone of Chinese language

The tone of Chinese language is like a musical scale; syllables with different pitch height have different meanings. The standard Chinese pronunciation have five formulations, they are YinPing (陰平), YangPing (陽平), ShangSheng (上声), QuSheng (去声) and QingSheng (軽声). The first four pronunciation are also called SiSheng (四声).

type of tone	*YinPing*	*YangPing*	*ShangSheng*	*QuSheng*
Name of tone	GaoPing (高平調)	GaoSheng (高昇調)	JiangSheng (降昇調)	QuanJiang (全降調)
Mark of tone	—	/	∨	\
Pitch height	5 5	3 5	2 1 4	5 1

(Explanation for Pronounce method)

Pinyin symbol	*Chinese characters*	*English word*
shī rén	詩人	a poet
shí rén	十人	ten persons
shǐ rén	使人	make (let) sb do
shì rén	世人	people
bái sè	白色	white
míng bai	明白	to understand

(Example for Pinyin symbol)

3 Chinese input system

3.1 General input method for Chinese

There are a lots of methods to input Chinese characters. General input methods are as follows.

- Standard PinYin input method.
- PinYin with tone mark input method.
- A style of input method with radical.
- GB (National standard in China) district code input method.

3.2 The problem of general input method

General input method is to translate from Pinyin to Chinese characters. For a foreigner Pinyin input method is probably the best one. If phonetic letters have several candidates for Chinese characters, users can add the tone, and the possible

numbers of Chinese characters are reduced a great deal. If users do not know the Pinyin, they can input Chinese character from its radical as is often used in looking up in a dictionary.

However, those methods are entirely "word by word" input. In this paper, we support a intelligent method to input "sentence by sentence" using statistical information between words.

4 Introduction of statistical information between words of Chinese

Using a dictionary is not very hard to change Pinyin to Chinese characters, The biggest problem through the input system using statistical information between words is the ambiguity.

For example, all the following sentences are made of homonyms (the same Chinese syllables but the included Kanji letters are different) without regarding tone. So one input syllable string may have several meanings.

Input: takantadezui.
Output:

她	看	她的	嘴	.
ta	kan	tade	zui	.

(She looks her mouse.)

她	看	他的	嘴	.
ta		tade	zui	.

(She looks his mouse.)

他	看	她的	嘴	.
ta	kan	tade		.

(He looks her mouse.)

他	看	他的	嘴	.
ta	kan	tade		.

(He looks his mouse.)

他	看	他	得	嘴
ta	kan	ta	de	zui

(* This is a wrong sentence.*)

In this paper we do not use any grammatical parsing, and MI between Chinese characters becomes the key to decide how the unity of syllable strings looks like a Chinese sentence.

5 MI and how to get the scores of sentences

5.1 Mutual information

A mutual information(MI)[1][2][3] is wide application today in various fields. It also can be used to see the relationship between two (or more) words in natural language processing (NLP).

The expression below shows the definition of the MI for NLP:

$$MI(w_1;w_2) = log\frac{P(w_1,w_2)}{P(w_1)P(w_2)} \quad (1)$$

w_i : a word
$P(w_i)$: the probability w_i appears in a corpus
$P(w_1,w_2)$: the probability w_1 and w_2 comes out together in a corpus

This expression means that when w_1 and w_2 have a strong association between them, $P(w_1)P(w_2) \ll P(w_1,w_2)$ i.e. $MI(w_1,w_2) \gg 0$. When w_1 and w_2 do not have any special association, $P(w_1)P(w_2) \approx P(w_1,w_2)$ i.e. $MI(w_1,w_2) \approx 0$. And when w_1 and w_2 come out together very rarely, $P(w_1)P(w_2) \gg P(w_1,w_2)$ i.e. $MI(w_1,w_2) \ll 0$.

5.2 How to get the score of a sentence

In this system, we input only Pinyin without tone, so one input syllable string can easily replaced into some 'Kanji sentences' with the given dictionary. However, to decide which sentences are correct is a difficult problem. The meaning of 'correct sentence' is a Chinese sentence which makes sense. For example, a syllable string "jingjidewenti" can bring up Kanji sentences like '経済的問題' (an economic problem), '経済得問題' and so on. The former answer makes sense, but the latter are not a correct Chinese sentences. The scores of Chinese sentences show how the given Chinese syllable strings look like Chinese. We use MI to select correct sentences from a lot of meaningless strings of Kanji characters.

Actually what we use in the calculation is not the real MI described in section 5.1. The MI definition in section 5.1 introduced the bigrams. A bigram is a possibility of having two certain words together in a corpus, as you see in the expression(1). Instead of the bigram we use a new possibility named d-bigram[3][4].

The idea of bigrams and trigrams are often used in the studies on NLP. A bigram is the information of the association between two certain words and a trigram is the information among three. A d-bigram is the possibility that two words w_1 and w_2 come out together at a distance of d words in a corpus.

For example, if we get 'Tom is a boy' as input sentence, we have six d-bigram data:

('Tom' 'is' 1) ('is' 'a' 1) ('a' 'boy' 1)
('Tom' 'a' 2) ('is' 'boy' 2)
('Tom' 'boy' 3)

('Tom' 'boy' 3) means the information of the association of the two words 'Tom' and 'boy' appear at the distance of 3 words in the corpus.

5.3 Calculation

The expression to calculate the scores using d-bigram between two words is[3]:

$$MI_d(w_1,w_2,d) = log\frac{P(w_1,w_2,d)}{P(w_1)P(w_2)} \quad (2)$$

w_i : a word
d : distance of the two words w_1 and w_2
$P(w_i)$: the possibility the word w_i appears in the corpus
$P(w_1,w_2,d)$: the possibility w_1 and w_2 come out d words away from each other in the corpus

The bigger the value of MI_d gets, the more those words have the association. And the score of a sentence is calculated with these MI_d data(expression(2)). The definition of the sentence score is[1]:

$$I_d(W) = \sum_{i=0}^{n}\sum_{d=1}^{m}\frac{MI_d(w_i,w_{i+d},d)}{d^2} \quad (3)$$

d : distance of the two words.
m : distance limit
n : the number of words in the sentence
W : a sentence
w_i : The i-th Kanji character in the sentence W

This expression(3) calculates the scores with the algorithm below:

1) Calculate MI_d of every pair of words included in the given sentence.

2) Give a certain weight according to the distance d to all those MI_d.

3) Sum up those $\frac{MI_d}{d^2}$. The sum is the score of the sentence.

Church and Hanks[1] said that the information between two remote words has less meaning in a sentence when it comes to the semantic analysis. According to the idea we put d^2 in the expression so that nearer pair can be more effective in calculating the score of the sentence.

6 The PCS system

6.1 Outline of PCS system

This system PCS supports the method to input only Pinyin without tone. And looking up in the dictionary, the system picks up the Kanji characters found in the given syllable strings. The system reads the syllable strings from left to right, to find out every possibility. Pinyin letters of the sentence are looked up in the dictionary and if they are found in the dictionary the system change them to corresponding Kanji characters. All the found Kanji characters are numbered by its position in the sentence.

After picking up all the Kanji characters in the sentence the system tries to put them together and makes them up to sentences.

Input Pinyin without tone.
⇓
Find out all possible syllable strings, and
translate them to corresponding Kanji characters.
⇓
Make up sentences with the kanji characters.
⇓
Calculate the score of sentences
using the mutual information.
⇓
Compare the scores of all the made-up sentences
and get the best-marked one
as the most 'Chinese-like' sentence.

Then the system compares those sentences made up with found Kanji characters and decides which one is the most 'Chinese-like'. For that purpose this system calculates the score of probability of each sentence (section 5.3).

6.2 The corpus

A corpus is a set of sentences. In this system, we need a corpus of Chinese sentences which are already segmented. We could not require the Chinese corpus of a large scale. So we had to prepare one for ourselves. We selected a raw Chinese text that seems suitable. The corpus we used in this paper has about 500 sentences. This corpus is too small as a model of the real world, however, the experimental results show that the system works efficiently even though the corpus is small.

6.3 The dictionary

The dictionary for PCS system is made of two parts. One is Pinyin letters without tone and the other is Kanji characters corresponded to Pinyin letters. There may be more than one Kanji character attached to one Pinyin. The second part which has Kanji characters is of type list, so that it can have several Kanji characters. This phenomenon appears properly in Chinese since input Pinyin without tone.

Chinese character : ("jin" ("金" "近" "今"))
Pinyin letters Kanji characters

Chinese phrase : ("jinnian" ("近年" "今年"))
Pinyin letters Kanji characters

7 Experiment

We described the structure for the corpus (section 6.2) and the dictionary (section 6.3) used in this system.

The dictionary and the statistical information used in this paper are got from the given corpus. So, the experimental result totally depends on the corpus.

The corpus used in this system have about 500 Chinese sentences which are already segmented. The dictionary have about 1500 words, some of which may not be in the corpus. The word contains sole Chinese character and Chinese phrase.

Here are some examples of implementing PCS for Chinese language (table 1) (table 2) as follows. The input is a string of Pinyin. Table 1 is the example for the best three and table 2 is the example for the best five answers. Table 2 is an ambiguous answer for homonym.

Table 1: Experiment in Chinese

Input: zhongxiwenruanjianjianrongzaiyiriqianlibushizhangai.

English meaning: There is no problem to use both Chinese and western software together.

correct answer	output sentences	scores
★	中西文 软件 兼容 再 一日 千里 不是 障碍 .	58.1717
	中西文 软件 兼容 买 一日 千里 不是 障碍 .	14.75879
	中西文 软件 兼容 在 一日 千里 不是 障碍 .	14.75879

Now we still used Japanese fonts istead of Chinese fonts, and we define the symbol " ü " is equivalent to " u " because we have no input method to input " ü " which is not a normal English character. So, the different pronunciation " nü " (女) from " nu " (努) is regarded as the same Pinyin syllable.

Table 2: Experiment in Chinese

Input: takantadezui.

English meaning: She (he) look her (his) mouth.

correct answer	output sentences	scores
★	她 看 她的 嘴 .	-16.92621
★	她 看 他的 嘴 .	-16.92621
★	他 看 她的 嘴 .	-17.09749
★	他 看 他的 嘴 .	-17.09749
	他 看 他的 最 .	-22.2115

8 Results

Putting all input Pinyin letters to test the system PCS, we get the score of each sentence made up of Kanji letters. After getting the list of sentences, we look for the most 'Chinese-like' sentence in the list. The data show the scores the 'correct' Chinese sentence got (table 3).

Table 3: Experiment for PCS

corpus	:	about 500 Chinese sentences (which are already segmented)
dictionary	:	about 1500 Chinese words (includes Chinese phrases and Chinese characters not in the corpus)
input	:	only Pinyin without tone
number of input sentence	:	about 100 each

	the best score	~ the second best	~ the third best
α	96.0 %	97.0 %	99.0 %
β	96.0 %	97.0 %	99.0 %
γ	94.0 %	96.0 %	98.0 %
δ	88.0 %	90.0 %	92.0 %

α : the same sentences in the Chinese corpus

β : the sentences one word replaced (the replaced word can be both in or not in the corpus)

γ : sentences not in the corpus (the words are all in the corpus)

δ : sentences not in the corpus (include the words not in the corpus)

According to the experimental results (table 3), it is obvious that the system PCS is very useful. The table 3 shows that most of the sentences, no matter whether the sentences are in the corpus or not, are changed into Chinese characters correctly. We find the most 'Chinese-like' sentence getting the best score in the list of possible sentences.

The Calculation of Quantitative Characteristics of Philological Dictionaries in UNILEX-D System

L.I. Kolodyazhnaya,
Russian Academy of Science,
Institute of Russian language;
Russian language Computer Fund
E-mail: irlras@irl.msk.su

Project note

AREA: computer lexicography

Summary:

The paper gives an account of the concept of the quantitative characteristic for the linguistic dictionaries of various genres and the methods of their calculation in the universal lexicographic processor UNILEX-D environment. The account is accompanied by some results, obtained in computer versions of Russian dictionaries.

1. Usually a dictionary shows explicitly only one quantitative feature - the number of items (head words) in a dictionary. From the point of view of the formal description of the dictionary item structure the head-word is only one of the item's components. Sometimes it is useful to know also some other quantitative features of the item's components which are important for the given dictionary.

For example, for the explanatory dictionaries it is important to obtain the distribution of meanings of the grammatical categories of words, the number of their lexical meanings, the number of style markers and so on.

The information about the distribution of the number of words entering the given number of the synonymic groups as well as the information about the size distribution of the synonymic groups is probably useful for synonym dictionaries.

The number of words, having variants or some other specific features can be useful for orthographicaldictionaries.

One can consider also more complicated quantitative characteristics which depend on more then one of the components of a dictionary item.

For example, for S.I. Ozegov's Dictionary of Russian is important to build the tables showing the dependence of the distribution of some category for the derivative word on that of the producing one.

2. The number of quantitative characteristics depends on the dictionary genre (explanatory, synonymous, bilingual, syntactic, grammatical, orthographical), and the structural complexity of dictionary item.

Though the structures of items for the dictionaries of various genres may differ from each other both in the components' type and their quantity, it is possible to calculate their quantitative characteristics by the samemethod.

One can consider also more complicated quantitative characteristics which depend on more then one of the components of a dictionary item.

The main principle of the universal lexicographic processor (or dictionary system) UNILEX-D is to work with the dictionary bases using the description of the dictionary item structure. This structure is described by the list of elementary and composite components, following the order which is adopted in the given dictionary.

Speaking in computational terms, UNILEX-D is some kind of lexicographic assembler which is characterised by the set universal transformations executed on the dictionaries and the schemes for dictionary data naming, similar to those for classical macroassemblers.

There are four types of transformations which are used by UNILEX-D to create new dictionaries from the old one : table, inversion, sampling and projection.

3. The total number of transformations which can be applied some dictionary depends on its item structure, namely on the number of the components in dictionary item. All possible transformations of some dictionary forms the so-called lexicographic space of this dictionary.

More precisely, lexicography processor makes possible to solve the next lexicography problems:

- to create a new item structure using the typical scheme;
- to input the information in the new dictionary's item, using the information of any other dictionary;
-to edit the information in this dictionary, using the information of any other source;
-to create the new dictionaries, using operations of sampling, projection, fusion and intersection;
-to create various inverted tables for the given dictionary and tables of quantitative distribution of meanings for any component of a dictionary item;
- to view simultaneously the information of two different dictionaries in the different windows of the screen.

It is possible not only to print the results on a printer, but to prepare the printed outlay of the dictionary for publishing, using the markes of the TeX-system.

Вычисление квантитативных характеристик филологических словарей на основе системы Унилекс-Д

Колодяжная Л.И.

Резюме:

В настоящем докладе излагается подход к описанию количественных характеристик лингвистических словарей различного рода и методам их получения с помощью универсального лексикографического процессора Унилекс-Д. Приводятся некоторые результаты, полученные на основе анализа компьютерных версий русских словарей.

Study of quantitative correlations between stylistics, grammar and polysemy of words (on the basis of Ozhegov's Dictionary)

L.I.Kolodjazhnaya,
Institute of Russian language
of the Russian Academy of Sciences,
Computer Fund of Russian Language,
Russia , 121019, Moscow,Volkhonka 18/2
E-mail: irlras@irl.msk.su.

A.A. Polikarpov,
Lomonosov Moscow State University
Department of Theoretical and Computational Linguistics,
Laboratory of General and Computer Lexicology and Lexicography;
Russia,117899, Moscow, Vorobjovy Gory,
1-st Building of Humanities, LGCLL (room 960),
E-mail: logos@logos.msu.su.

Topical paper

AREA: General, Computer and Quantitative Lexicology

Summary:

The paper contains the results obtained on the basis of analysis of the database of "Russian Language Dictionary" (by S.I. Ozhegov, 22-th ed., Moscow, 1990) analysis. The purpose is to reveal quantitative correlations between stylistic, grammatic and polysemy features of words in a dictionary of the so called "short" type.

1. The present paper deals with some quantitative regularities of interrelations between stylistic characteristics of words and their quantitative-polysemic and grammatical (part-of-speech) characteristics obtained from "Russian Language Dictionary" by S.I. Ozhegov [Ozhegov, 1991] (further - RLD). The combination of these three characteristics allows to touch a very important knot of interrelations of external-functional and internal-systemic aspects in the vocabulary organization. The external-functional aspect involves stylistic characteristics that, in a special way, characterize the functional rangeand the effects in the use of lexical units. The internal-systemic aspect in this case is represented by quantitative-polysemic and part-of-speech characteristics of words. The number of meanings is a manifestation of the words' status, proximity to the core of language [Polikarpov, 1987]. Part-of-speech status represents the most general, categorial characterization of the words' semantics. Each categorial and polysemic group of words is also oriented in some way towards the centre/peripheral relations in language, is characterized differently in terms of age, frequency, functional area of use (stylistics). [Polikarpov, 1994].

Using a representative lexical source like RLD allows to pose a number of both theoretical and practical questions.

2. Various stylistic markers employed in the dictionary were reduced to four main categories which allowed to classify all units of the lexico-stylistic stock of the dictionary (meanings stylistically marked) into four respective lexico-stylistic classes: "bookish-specilized" (B), "colloquial" (C), "obsolete" (O), and "regional" (R).

3. One of the most important parameters of lexical units is their polysemy,in some way presenting words' semantic "size". This characteristic specifically reflects the scope of functional potential of the word, the width of its sense reference. This general semantic feature of words must be closely related to different qualitative characteristics of their meanings, e.g., to their stylistic quality: the greater the word polysemy, the wider is referential area of each word meaning, the weaker is they bound to some specific functional area, the lower is their degree of stylistic marking. This is why stylistical marking of words meanings will be considered separately in different polysemic zones.

4. All lexical units of the dictionary are divided into following polysemic groups:

1) words having 1 meaning,
2) words having 2 meanings,
3) words having 3 to 4 meanings,
4) words having 5 to 8 meanings,
5) words having 9 to 16 meanings,
6) words having 17 to 32 meanings, etc.

5. We use part-of-speech categorization as follows: nouns, verbs, adjectives, adverbs, pronouns, prepositions, particles, conjunctions, interjections, parenthetic words, predicates. We consider nouns as having the most concrete categorial semantics oriented in most cases on denoting real physical things, but not their features and relations. Verbs and adjectives are more related to denoting features of things, relations and ideas. Pronouns, conjunctions and prepositions even in the more explicit manner are oriented in this direction.

6. The study was conducted using the database of RLD created in the Computer Fund of Russian Language. UNILEX-D program was used as the instrument for selecting and counting differently styllistically marked meanings of different parts-of-speech words belonging to different polysemy zones.

7. The most important consideration for selecting the object for our investigation was the fact that the "short"-size dictionary should fully enough reflect the i n t e r s e c t i o n of individual normative active vocabularies of speakers of some language at some certain time. I.e. it should provide some approximate description of the core of the total active normative lexical stock of a given language community in a

α is the data when the input sentences are all the same in corpus. But as is shown in table 2, the first four are all make sense, so we can not get perfect result. This is a problem for Chinese homonym.

PCS does not check the corpus itself when it calculate the score. It just use the MI_d, the essential information of the corpus. That is, whether the input sentence is written in the corpus or not does not make any effect in calculating scores directly. However, since PCS uses MI_d to calculate the scores, the fact that every two Chinese words in the sentence have connection between them raises the score higher.

When the input sentences are not in corpus, the ratio of correct answer gets down (see table 3, data δ).

MI_d comes to be the key to use the effect of the mutual information between Chinese words indirectly so that we can put the information of the association between Chinese words to practical use. This is what we expected and PCS works successfully at this score.

9 Conclusion

This paper shows that this input system PCS is quite intelligent and efficient for translate syllables into Chinese characters without using any grammatical information. Instead, this experimental system PCS makes it possible to input Pinyin "sentence by sentence", and using mutual information between Kanji words we can choose the most 'Chinese-like' sentence from all the possibilities. According to the results of the experiments, PCS can change almost all the Pinyin sentences to Chinese characters sentence 'correctly'. This result is considerably good enough.

The result shows that using MI_d between Chinese words is a very effective method for Chinese input system.

10 Future works

The best corpus for PCS is the one which has enough Chinese sentences which can get the correct statistical information. Scaling up the corpus is one of the biggest problems. The corpus we used was a small one, however, even with that small corpus we could see that PCS works efficiently. When it comes to the corpus which has much more grammatically illegal sentences, we can not say that this system works effectively. However, the result of this paper is hopefully good enough.

Using this method, we look forward to putting this system to practical use.

References

[1] Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. Parsing, Word Associations and Typical Predicate-Argument Relations. *International Parsing Workshop*, 1989.

[2] Frank Smadja. How to compile a bilingual collocational lexicon automatically. *Statistically-based Natural Language Programming Techniques*, pages 57–63, 1992.

[3] Junya Tsutsumi, Tomoaki Nitta, Kotaro Ono, and Shiho Nobesawa. A Multi-Lingual Translation System Based on A Statistical Model (written in Japanese). *JSAI Technical report, SIG-PPAI-9302-2*, pages 7–12, 1993.

[4] Shiho Nobesawa, Junya Tsutsumi, Tomoaki Nitta, Kotaro Ono, Sun Da Jinag, and Masakazu Nakanishi. Segmenting a Sentence into Morphemes Using Statistic Information between Words. *COLING94*, 1994 (to appear).

[5] David M.Magerman and Mitchell P.Marcus. Parsing a Natural Language Using Mutual Information Statistics. *AAAI*, 1990.

[6] P.Brown, J.Cocke, S.Della Pietra, V.Della Pietra, F.Jelinek, R.Mercer, and P.Roossin. A Statistical Approach to Language Translation. *Proc. of COLING-88*, pages 71–76, 1989.

Table 2

Distribution of stylistically marked meanings of different parts of speech in different polysemic zones of RLD.

Polys zones	All (abs.) words	All (abs.) meanings	Stylistically marked meanings in per cents as related to all (marked and unmarked) meanings B %	C %	O %	R %	All %
NOUNS							
1	13299	13299	10,90	11,35	4,49	0,54	27,28
2	3982	7964	8,55	8,43	3,92	0,35	21,25
3-4	1321	4393	9,15	7,42	3,19	0,25	20,01
5-8	213	1176	9,27	6,29	2,81	0,26	18,63
9-16	14	138	10,90	8,70	1,40	-	21,00
1-16	18829	26970	9,88	9,61	2,91	0,42	22,82
VERBS							
1	4976	4976	10,87	37,38	5,22	0,26	52,17
2	2191	4382	4,47	20,6	2,26	0,09	27,22
3-4	964	3180	2,77	20,82	1,92	0,06	25,57
5-8	236	1343	1,64	12,58	0,74	0,07	15,03
9-16	36	458	2,62	17,47	1,1	-	18,57
17-32	4	77	1,30	5,19	-	-	6,49
1-32	8407	14416	5,96	25,48	3,02	0,13	34,59
PREDICATES							
1	149	149	0,67	59,06	2,68	-	62,41
2	23	46	-	21,74	-	-	21,74
3-4	3	9	-	44,44	-	-	44,44
1-4	175	204	0,49	50,00	1,96	-	52,45
ADJECTIVES							
1	3824	3824	12,81	11,17	4,71	0,10	28,79
2	1448	2896	6,28	6,56	2,52	-	15,36
3-4	520	1696	3,66	7,13	2,06	0,18	12,85
5-8	91	527	2,66	8,73	1,14	-	12,53
9-16	11	113	1,80	10,60	7,10	-	19,50
1-16	5894	9056	8,28	8,79	3,33	0,05	20,45
ADVERBS							
1	681	681	4,40	32,30	4,84	0,15	41,69
2	95	190	1,58	23,68	3,16	-	28,42
3-4	25	79	-	24,05	-	-	24,05
5-8	5	25	-	36,00	-	-	36,00
1-8	806	975	3,38	30,05	4,00	0,10	37,53
PRONOUNS, CONJUNCTIONS, PREPOSITIONS							
1	238	238	2,10	14,29	8,82	0,84	26,05
2	38	76	-	25,00	1,31	2,63	28,94
3-4	39	131	-	9,16	1,53	-	10,69
5-8	17	97	-	4,12	-	-	4,12
9-16	5	52	-	3,84	-	-	3,84
1-16	337	594	0,84	11,61	4,04	0,67	17,16
PARTICLES, INTERJECTIONS, PARENTHETIC WORDS							
1	273	273	1,56	31,66	3,13	-	36,35
2	39	78	-	44,86	-	1,28	46,14
3-4	10	33	-	30,30	-	-	30,30
5-8	2	12	-	50,00	-	-	50,00
1-8	324	396	1,01	33,33	2,02	0,23	36,59
ALL PARTS OF SPEECH TOGETHER							
1	23440	23440	10,75	17,99	5,90	0,39	35,03
2	7816	15632	6,79	11,98	3,14	0,22	22,13
3-4	2882	9521	5,79	12,12	2,49	0,13	20,53
5-8	564	3180	4,53	9,68	1,54	0,12	15,87
9-16	66	761	3,67	13,14	1,83	-	18,64
17-34	4	77	1,29	7,79	1,29	-	10,37
1-34	34772	52611	8,21	14,55	3,03	0,27	26,06

Here: B - bookish, C - colloquial, O - obsolete, R - regional.

Hurst's Law as a Universal Law of Quantative Linguistics of a Coherent Text

Y.K. Krylov,
St. Petersburg Electrotechnical University
the Department of High Mathematics,
E-mail: polikarp@logos.msu.su

Topical paper

AREA: Stochastic processes in language

Summary:

Relevance of Hurst's Law is discussed in respect to the statistic structure of a coherent text.

1. Let

$$x(1), x(2),... x(t),... x(n),... x(N) \qquad (1)$$

be a succession of observed values of a system of random variables (X(1), X(2),..X(t),...X(n), ..X(N), the substantial interpretation of which is irrelevant for us yet. Later on, according to the accepted terminology, we will call (1) a "time series" or a trajectory of a stochastic process X. Let us now consider a segment of n primary numbers of the series (1).

Let us designate by $< x(n) >$ their mean value:

$$< x(n) >= \frac{1}{n} \sum_{t=1}^{t=n} x(t) \qquad (2)$$

$$< S(n) >= \frac{1}{n-1} \sum_{t=1}^{t=n} (x(t) - < x(n) >)^2 , \qquad (3)$$

- a nondisplaced estimation of the dispersion calculated on the baseof this segment.

Let us take

$$Z(t,n) = \sum_{t=1}^{t=n} (x(t) - < x(n) > = Z(t)-t*< x(n) > \qquad (4)$$

- a deviation of t-partial sum of the series (1) from average $<x(n)>$ which was accumulated by the "moment" t. The difference of maximum and minimum values of Z (t,n) (t=1,2,...n) will be called a "swing" R(n). Therefore, by definition,

$$R(n)=\max Z(t,n)-\min Z(t,n) \qquad (t = 1,2,....n) \qquad (5)$$

Further, we will switch to the unmeasurable relation r(n) = R(n)/S(n) which we are going to call a "standard swing". Using this unmeasurable relation it will be possible to compare swings for different distributions.

2. The method of analysis of stochastic successions, exposed above, is called the "Method of R/S-Analysis". It was offered by H.E. Hurst and described in detail in his monograph [1]. Obviously, a "standardized swing" is to depend on n - number of members of the series (1) which are used for its calculation. It is shown in [2,3], that for a purely random time series the equally distributed stochastic quantities possessing the finite dispersion:

$$r(n) = SQRT (Pi*n/2), \qquad (6)$$

where SQRT - an operation of square root extraction, Pi = 3.14157. I.e. it is described by the power-dependance with an exponent which equals 0,5.

Having analysed many nature processes such as a flow of rivers, a level of precipitations, a quantity of flaky deposits, a size of rings and an index of tree remification, Hurst [1] established, that a standardized swing, observed by him, can be described sufficiently well by the empirical formula:

$$r(n)=(n/2)^H, \qquad (7)$$

i.e. in bylogarithmic coordinates experimental points lay on a straigt line with a sufficiently high exactness.

$$\ln r(n)=H*(\ln n - \ln 2) \qquad (8)$$

Along with that, it also occured, that for different phenomena Hurst's exponent H is more or less symmetrically distributed around the average value 0,73 with a standard deviation which approximately equals 0,09. Hurst's observation, exposed above (that for many natural processes r(n) can be described by the power-dependance with the power exponent H, which significantly exceeds 0,5) was called "Hurst's Law" [4].

3. In the offered work the realization of Hurst's law for time series, given in the texts of natural languages, was analysed. Let us designate by

$$g(1), g(2),.... g(t),..... g(n),....g(N) \qquad (1b)$$

a succession of word occurences which are realised in a given text. Then it will be possible to designate by $x(t) = x (g(t))$ any quantitiative characteristics of some certain word. Obviously, while a coherent text generation all its elementary fragments are statistically dependent, and, if for a considered series the standardized swing r(n) is really described by the power-dependance (7), then Hurst's exponent H may serve as an integral measure of the interconnection of the context in relation to its X characterisation. The realizability of Hurst's law has been checked for the series such as (1), as well, as for the successions of numbers taken out of (1), with the help of corresponding transformation of the initial series. For series which are directly given, as X, were such characteristics are concidered as length of a word (calculated in graphemes, syllables, or morphemes); potential or realised in texts polysemy; coverage of a text by words of its vocabulary (in the latter case $X(t) = 1 / F(t)$, where F(t) - absolute frequency of a word with a successive number t in this text); distribution of words of a fixed class W in a text ($x(t) = 1$, if $g(t)$ belongs to W and $x(t) = 0$, when $g(t)$ does not belong to W). Here by W we designate concrete lexemes; once,

twice, F times used words in texts; syntactic characteristics of a context (when x(t) = 1 before some definite punctuation mark - in the end of a sentence, paragraph, etc.).

When analysing distributions of words of class W in text, (1) is a succession of ones and zeroes. In that case, together with the initial series, the series of ordinal numbers (addresses) of the undersuccession of units of initial series were also considered. Along with that, the first differences of addresses had a sense of "distances" between successive appearances of words of class W in text and also were investigated by methods of R/S-analysis.

For very long texts, in order to abbreviate the process of calculation, together with (1) the series of groupped data were also used. In the latter case the initial series was devided into intervals, with a fixed step h and in every interval the sum of values x(t) was calculated (e.g., a number of appearances of a fixed lexeme in non-crossing each other segments of texts of length 100 or 1000 word occurences). Later on, for further analysis, we will take a series, which was obtained as a result of the groupping.

Finally, besides the study of trajectories which correspond to certain texts for some quantitiative characteristics there were built series representative enough also in the intertext area. In this series of experiment, by n we designate the length of some whole texts, for each of them there was calculated a standardized swing r(n). Using different texts, it was possible to check the realization of Hurst's law for the case of increase of text length.

4. The investigations showed, that Hurst's law (the power-dependance (7) with Hurst's exponent, which significantly exceeds 0,5) really was relevant for the substantial majority of the investigated time series. With the average value 3/4 the average square deviation of Hurst's exponent turned out to be much less than 0,1. Moreover, these results made it possible to bring up a hypothesis, that at least for the texts of approximately the same length the parametres of power dependance (7) keep constant values for the time series of different linguistic objects. Thus, e.g. (as in many other experiments), experimental points for the covering by once-used words of the of the story "Malva", by M. Gorky, and the distributions of the sentences of different lengths in the story "King Lear from the steppe", by I. Turgenev, practically layed on the same straight line (in the system of bylogarithmic coordinates).

The check-up of the hypothesis, stated above, which is based on the fact, that Hurst's exponent really is a universal constant for texts of some natural language, needs a realization of a very difficult calculative experiment, which is in progress nowadays. If this hypothesis will have a good empirical substantion, we have a true evidence of the fact, that the production of the text is not connected with any major level of its organization, but takes piace simultaneously on all levels, which cause each other, that may testify the fractal nature of the considered phenomenon.

References

1. Hurst H.E., Black R.P., Simaika Y.M. Long-Term Storage: An Experimental Study. - London, 1951.
2. Feller W. The Asymptotic Distribution of The Range of Sums of Independent Variables // Ann. Math. Stat. 22, 427-432; 1951.
3. Hurst H.E. Long-Term Storage Capacity of Reservoirs // Trans. Am. Soc. Civ. Eng. 116, 770-808, 1951.
4. Feder E. Fractals. M., 1991.

Закон Херста как универсальный закон квантитативной лингвистики связного текста

Крылов Ю. К.

Резюме:
Рассматривается выполнимость закона Херста на статистических структурах связного текста.

Lemmatizator as the Recognizing Word Model

Valentina Krytskaya,
O.O.Potebnya Institute of Linguistics
of the Academy of Sciences of Ukraine
Hrushevsky Street 4, Kiev 1, 252001, Ukraine

Phone: (044) 228-2680

Project note

AREA: Automatic lemmatization

SUMMARY:
The test of the scientific hypothesis of utilization of the formalized inflective characteristics of words for typological analysis of Russian and Ukrainian, and of getting the computer results of the lemmatization.

The lemmatizator is a part of the computer grammar. The article proposes an approach to the procedure of lemmatization as an algorythmic model of transformation of word-forms of the word into one form. An attempt is made to determine the characteristics of languages (Russian and Ukrainian) on the base of analysis of lemmatizator parameters.

Лемматизатор как модель распознавания слова

Критская В.

РЕЗЮМЕ:
Проверка научной гипотезы использования формализованных изменяемых характеристик слов для типологического анализа русского и украинского языков и получения компьютерных результатов лемматизации.

increases. It's important to note, however, that a mean frequency of more generalized unit is not only higher, then that of unit, which is its variant, but it grows more rapidly on increasing segments of text. Mean frequency of hyperlexemes usage increases with growth of text size more rapidly, then that of lexemes, which, in turn, grows quicker, then mean frequency of word-forms. That is caused mainly by the fact, that for more and more frequent hyperlexemes a wordforming and wordchanging potential continues to realize. One can imagine a situation, when while corpus growth the lexemes dictionary growth stops, while growth of word-forms dictionary continues.

5. As a result of ordering word-forms, lexemes, hyperlexemes by absolute frequency decreasing, we got the rank-frequency distributions, which reflect the functional dependence between the absolute frequency of lexical units (Fi) and their rank (i) (see Fig.1).

As one can see, the grafic presentation of the rank-frequency distribution indicates Zipf-Mandelbrot's law parameter "gamma" different values for different lexical units. For lexical units, correlated in the feature "variability - invariability" this parameter corresponds to relative degree of their rank-frequency distribution heterogeneity, higher degree of word-forms specificity as compared to lexemes and the same to lexemes as compared to hyperlexemes.

6. Such integral index as word-forms mean polytexty and lexemes mean polytexty (in 7 topic chapters of the corpus) demonstrates also higher degree of specificity of word-forms than that of lexemes:

$\bar{P}$ (V) = 2,04 $\bar{P}$ (L) = 2,41.

Fig.1
Rank-frequency distribution of word-forms, lexemes, hyperlexemes (bilogarithmic scale) in the corpus of legal acts

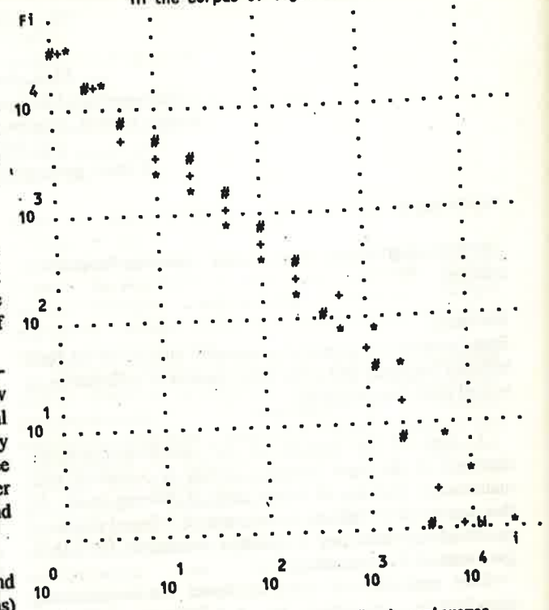


Table 1

Chr		F1	F2	F5	F6	F7	F8	F9	Fs	P	Vl	Vs
ZAKON		162	46	17	86	116	50	51	528	7	3	17
ZAKON	n	151	46	14	59	112	49	48	479	7		10
zakon		20	1		1	1	1	2	26	6		
zakona		13	4		1	5	2	8	33	6		
zakonam		1	7			4			12	3		
zakonami					4	3	1	6	14	4		
zakonah			1		1				2	2		
zakone		7							7	1		
zakonov		19	13	9	31	89	40	29	230	7		
zakonom		56	17	5	20	6	3	3	110	7		
zakonu		34	3		1		2		40	4		
zakony		1				4			5	2		
ZAKONNOST	n	2		2	5	4	1	2	16	6		2
zakonnosti		2		2	4	1		2	11	5		
zakonnost					1	3	1		5	3		
ZAKONNYJ	a	9		1	22			1	33	4		5
zakonnuju		6							6	1		
zakonnyje		1			2				3	2		
zakonnym		1			1				2	2		
zakonnymi					1				1	1		
zakonnyh		1		1	18			1	21	4		

Table 2

	N	V	L	V/N	N/V	L/N	N/L	V/L
Ch. 2	12078	2224	1205	0,184	5,431	0,100	10,023	1,846
Ch. 5	16477	3179	1687	0,193	5,183	0,102	9,767	1,884
Ch. 9	40550	5711	2806	0,141	6,443	0,069	14,451	2,035
Ch. 1	44792	6294	3040	0,141	7,117	0,068	14,734	2,070
Ch. 6	58266	8437	4114	0,145	6,901	0,071	14,163	2,051
Ch. 8	74858	8652	3945	0,116	8,652	0,053	18,975	2,193
Ch. 7	128539	10712	4595	0,083	11,999	0,036	27,974	2,331
Over. corp.	375560	22121	8693	0,059	16,98	0,023	43,20	2,545

References

1. Boroda M.G., Polikarpov A.A. (1984). Zakon Zipfa-Mandelbrota i edinicy razlichnyh urovnej organizacii teksta (Zipf-Mandelbrot's Law and various levels of text organization units). In: Kvantitativnaja lingvistika i avtomaticheskij analiz tekstov. Acta et Commentationes Universitatis Tartuensis, issue 609. Tartu: Tartu University Press.

2. Koehler R. (1986). Zur linguistischen Synergetik: Struktur und Dynamik der Lexik. - Bochum.

3. Polikarpov A.A. (1988). Logicheskoje prostranstvo edinic leksicheskoj podsistemy jazyka i kvantifikacija otnoshenij mezdu nimi (Logical space of lexical subsystem units and quantification of their correlations). In: Papers from the Scientific Conference on Applied Linguistics and Automatic Text Analysis.- Tartu.

Квантитативный анализ распределений лексических единиц различных уровней в юридических текстах: словоформы, лексемы, гиперлексемы

Москаленко Т.А.

Резюме:
Рассматриваются некоторые проблемы квантитативно-системного анализа отношений между частотой и некоторыми другими характеристиками лексических единиц различных уровней.

The Inflexion of Nouns in the PC Aspect

Orlova Larisa
Kiev, O.O. Potebnya Institute of Linguistics
of the Academy of Sciences of Ukraine
Hrushevsky Street 4, Kiev 1, 252001, Ukraine
Phone: 2282680

Projct note

AREA: Automatic morphological analusis.

Summary:
The test of the scientific hypothesis of utilization of the formalized inflective characteristics of the noun for typological analysis of Russian and Ukrainian, and of getting the computer results of the paradigm synthesis.

The work offers the computer approach to the construction of models for the substantival inflexion. This approach is based on the utilization of combinations of the formally selected parametres, which consider differential indications of paradigms. The accepted approach may be applied to the automatical synthesizing of a paradigm, and it can serve the objective grounds for comparative analysis of the inflexion in inflexional languages.

Флективное изменение имен существительных в компьютерном аспекте

Л. Орлова

Резюме:
Проверка научной гипотезы использования определенных изменяемых характеристик существительного для типологического анализа русского и украинского языков и получения компьютерных результатов синтеза парадигм.

В поисках синергетических механизмов языка

Р.Г. Пиотровский
Санкт-Петербургский Педагогический Университет

Доклад

ТЕМАТИЧЕСКАЯ ОБЛАСТЬ: Теоретическая лингвистика

Резюме:
Определяется круг проблем, которые должны быть решены для вскрытия природы синергетических механизмов языка и речи.

Несмотря на искусное сочетание идей алгебраической лингвистики, информационно-вероятностных подходов, а также теории нечетких множеств и лингвистической переменной, современные системы автоматической переработки текста обладают пока очень слабыми способностями к самоконтролю и самоорганизации. Причина состоит в том, что мы почти ничего не знаем о тех синергетических механизмах, которые управляют нормальным функционированием системы языка и речи. И это не случайно. Система общенародного языка и речемыслительная деятельность отдельного человека функционируют в норме удивительно слитно и слаженно, наглухо закрывая те "окна", через которые можно было бы наблюдать лингвистическую синергетику.

Возможность наблюдать синергетические механизмы языка и индивидуальной речи предоставляют нам:

во-первых, ситуации лингвистических катастроф, т.е. интенсивное смешивание языков, приводящее к креолизации и пиджинизации языка-победителя и разрушению побежденного языка, а также патология речемыслительной деятельности человека, возникающая при измененных состояниях сознания, при локальных поражениях мозга и эндогенных заболеваниях,

во-вторых, анализ лексико-грамматических ошибок и синтаксико-семантических нарушений в текстах, выдаваемых компьютером.

В ходе этих исследований лингвостатистика и информационная методика призваны выполнить две задачи:

во-первых, выявить значимые отклонения от нормы языка и речи в разрушенных и патологических текстах,

во-вторых, выявить типичные ошибки и "поломки" в системе языка и речи,

в-третьих, фиксировать значимые отклонения в распределениях лингвистических единиц (включая отклонения в параметрах закона Ципфа),

в-четвертых, описать информационные особенности разрушенных и патологических текстов.

Опираясь на все эти данные, можно будет строить гипотезы о природе синергетических механизмов, обеспечивающих сохранность и нормальное функционирование языка и речи.

In Search of Synergetic Mechanizms of Language

Piotrovsky R.G.

Summary:
It is consideres a range of questions, which should be put for revealing synergytic mechanizms of language and speech.

База данных синонимов русского языка и ее квантитативно-системный анализ

Покровская Е.А.
Московский государственный университет
Филологический факультет,
Лаборатория общей и компьютерной лексикологии и лексикографии,
Россия, 117899, Москва, Воробьевы горы, МГУ,
1 гуманитарный корпус, ЛОКЛЛ (комн. 960)
E-mail:polikarp@logos.msu.su

Topical paper

AREA: Quantitative Sustem Lexicology

Резюме:
Настоящая работа посвящена исследованию синонимических отношений в лексике в системном аспекте. На материале "Словаря синонимов" под редакцией А.П. Евгеньевой создана база данных из 500 двучленных и 300 многочленных групп. Анализируется зависимость объема синонимической группы, степени синонимической активности слов, чьи лексико-семантические варианты (ЛСВ) входят в синонимические отношения, от частиречных, полисемических, стилистических, фразеологических, деривационных характеристик и возраста слов.

Настоящая работа посвящена исследованию синонимических отношений в лексике в системном аспекте. Из "Словаря синонимов" под редакцией А. П. Евгеньевой была сделана выборка синонимических групп с минимальным и максимальным количеством членов, соответственно 500 двучленных и 300 многочленных групп (начиная с 8-членных). Каждый синоним-ЛСВ описывается по 8 параметрам, а именно: часть речи; стилистическая характеристика ЛСВ; число значений всего слова; стилистические характеристики всех ЛСВ в составе слова; возраст слова; наличие у слова фразеологически связанных значений; словообразовательный статус слова; является ли слово заимствованным.

В работе исследуется зависимость таких характеристик, как объем синонимической группы, в которую входит слово данным ЛСВ, способность слова вступать в синонимические отношения с различным числом слов от вышеперечисленных параметров.

В качестве источника эмпирических данных использовался "Словарь современного русского литературного языка" [первое издание - тт. 1-17, 1948-1965; второе издание, тт. 1-4, 1991-1993], "Словообразовательный словарь русского языка" А.Н.Тихонова, а также ряд этимологических словарей.

Были получены следующие результаты:

Часть речи и объем синонимической группы: наиболее активно в синонимические отношения вступают глаголы, прилагательные и наречия; существительные склонны в большей степени к образованию двучленных, а не многочленных групп.

Стилистические характеристики: стилистически окрашенные синонимы характерны для многочленных синонимических групп; в двучленных группах стилистически маркированные синонимы встречаются намного реже; наибольшую долю среди всех стилистически маркированных ЛСВ составляют единицы с пометами "разг." или "прост."; ЛСВ с пометой "книжно-специальное" концентрируются в двучленных синонимических группах.

Полисемия: чем выше полисемия слова, тем выше в среднем степень синонимической активности каждого его значения;

Возраст: в среднем степень синонимической активности значений слова возрастает с увеличением его возраста.

Фразеология: слова, имеющие фразеологически связанные значения, концентрируются в многочленных синонимических группах;

Деривационные характеристики: высокообъемные синонимические группы характеризуются большим количеством производных слов.

Database on Russian Synonyms and its Quantitative-Systemic Investigation

Elena Pokrovskaya

Summary:
The present paper is concerned with quantitative-systemic investigation of synonymic relations in lexicon in quantitative aspect. A database of 500 groups consisting of 2 synonyms and 300 groups consisting of 8 and more synonyms was created on the basis of "Dictionary of Synonyms" by A. Yevgenieva. It is analysed the dependence of volume of synonymic group, degree of activity of words which lexico-semantic variants are engaged in synonymic relations, on such features as part of speech, polysemic, stylictic, derivative characterictics, and age of a word.

Statistic Simulation of Text With Reference to Dynamics of its Units

Dobrina Rajnova,
Sofia University, Bulgaria/Moscow State University

Topical paper

AREA: statistic study of different language levels lexical units

Summary:

In this paper a new model for statistical modelling of texts is introduced with reference to some main parts of speech - nouns and verbs. This model includes a new-type rank-list - motivational frequency count - where word-forms of lexemes with only one motivation are combined in one general structure. The rank-position of a model is determined as the sum of frequencies of the word-forms that constitute it. The motivational frequency count is related to a distributional list with reference to all the contexts where a certain word-form in a certain knot appears.

Statistic simulation of texts is usually construed as sort of a lexico-statistic analysis based on representative samples that results in compiling rank or rank-distributive lists of word-forms and lexemes.

These lists are then used to characterize the peculiarities of style, sublanguages, genres, etc.

The modelling is usually carried out at the word level. For this end a very simple definition of a word is used, where a word is any stretch of text between two blank spaces. 1)

Such a simplified approach results in a number of word counts reflecting frequencies of occurrence of semilexical-semigrammatical units.

It is evident that a lexico-grammatical unit such as a word-form turn out of its context does not provide precise information on either the vocabulary or grammar of the language.

It is hardly more profitable to use a list of lexemes instead of a word-form list because the word-forms constituting the paradigm of a lexeme for reasons of statistic analysis lose information on the meanings (lexico-statistic variants) that they had in the text. But, on the other hand, in quantitative linguistics nowadays a number of similar rank lists of word-forms and lexemes have been compiled, different in length, where the first rank region of highest frequency items consists of relative, linking elements of text along with an insignificant quantity of nouns and verbs.

In their article 2) I. Sh. Nadareishvili and J.K. Orlov make a step further. They suggest to present the results of statistics of text not by one but three lists.

1. Alphabetical list with reference to frequencies and numbers of particular positions of words in increasing order.

2. Word list with reference to word frequencies and their first appearance in the text.

3. Word frequency could where words are entered in the increasing order of their frequencies.

From the point of view of these authors such lists provide exhaustive information about the text. It would be so if the text were a stochastic stationary process and the monosemous word were it's only unit. Since it is not the case, even such detailed lists bear very scarce information about the semantics and structure of the text and its units.

We need in linguistics to elaborate different semantic frequency dictionaries reflecting relative significance of various, mainly even unexplored yet, linguistic pecularities.

In our opinion, any statistic analysis has to reflect the semantic frequencies and distributions of such units that designate or can designate objects. These units are autosemantic, but not linking words which only indicate intratextual relationships. The frequencies and distribution of main words in a text (primary nouns and verbs) should be described in connection with their derivatives that possess motivation.

The statistic analysis should aim farther than any particular lexeme which has to be considered as an element of a major entity where actual paradigmatic system of a language is manifested. The main system-forming feature of a vocabulary is the intralinguistic motivation of lexical units 3). On the basis of their inner form and motivation lexical units can be combined in motivational nodes 4). Only 2-3% of all words in the system are not motivated (singular words) 5).

On this basis we can determine a given lexical unit as belonging to some word-building node, to a certain part of speech (general functional aspect) and to some type of reference (onomasiological aspect), since "a word is not an equivalent of a perceptible object but of how it is interpreted in a speech-generating act in the very moment of its coinage" 6). Thanks to motivational nodes the living inner form of a particular word is discovered and the relationships of its semantics and structure.

To us the motivational node is very close to what A.A. Polykarpov has defined as hyperlexeme 7). The systematic character of the vocabulary for a native speaker appears as a sum of various hyperlexemes and rules of usage thereof derived from speech.

In this research we included 2 representative sets of texts, one set contained 100 000 running words from newspapers and the other 200 000 running words from 'Stenographic Diary of the People's Assembly'.

In the first rank region of the count there are 29 primary nouns representing motivational families in the 29 first positions of the count. In the second rank region (frequency 275 to 4) there are 912 entries, 95 entries with frequency 4, 106 entries with frequency 3, 265 entries with frequency 2, 557 entries with frequency 1 (0.27% of all primary nouns).

In the paper we also give an example of how the statistic analysis works on the second stage of research where the motivational list combines with the frequency count of distribution of lexical units - lexemes and word-forms 9).

References

1) G. Glison. Vvedenie v deskriptivnuju lingvistiku. Moskva, 1959.

2) I.S. Nadarejshvili, J.K. Orlov. Metod polnoj fiksatsii teksta pri lingvostatisticheskom analize // Linguistica XII. - Tartu, 1978.

3) O.P. Ermakova, E.A. Zemskaja. Sopostavitel'noe izuchenie slovoobrazovanija i vnutrennjaja forma slova // Izvestija AN SSSR, serija literatury i jazyka, tom LXIII. - M., 1985.

4) D. Rajnova. Strukturno-semanticheskoe opisanie leksicheskih sistem // Aktual'nye problemy russkogo slovoobrazo-vanija. Uchenye zapiski, t. I. - Tashkent, p. 453.

5) D.A. Rajnova, A.I. Kuznetsova. Struktura i raspredelenie slov-odinochek v bolgarskom i russkom jazykah // Ispol'zovanie matematicheskih modelej i EVM v lingvistike. Sofija, BAN, 1976, p. 295-299.

6) W. Gumbol'dt. Izbrannye trudy po jazykoznaniju. M., 1984.

7) A.A. Polikarpov. Logicheskoe prostranstvo edinits leksicheskoj podsistemy jazyka i kvantifikatsija sootnoshenij mezhdu nimi (Logical Space of Lexical Units Subsystem of Language and Quantification of Relations between them) // Prikladnaja lingvistika i avtomaticheskij analiz teksta. - Tartu, 1988, p. 67-70; G.O. Karimova, A.A.Polikarpov. Printsipy vydelenija giperleksemy kak edinitsy leksicheskoj podsistemy (Principles of Defyning Hyperlexem as a Unit of Lexical Subsystem) // Derivatsionnye tipy i gnezda v sinhronii i diahronii. - Vladivostok, 1989, p. 158.

8) D.A. Rainova. Nekotorye problemy lingvisticheskoj verojatnosti i valentnosti. KD, Moskva, MGU, 1980, p. 107 and next; D. Rajnova. O statisticheskom analize i sinteze teksta. - Sofija, 1994 (in press).

Статистическое моделирование текста в аспекте динамики его единиц

Райнова Д.

Резюме:

В настоящем докладе представлена модель статистического моделирования текстов в отношении главных частей речи - имен существительных и глаголов. Эта модель включает рангово-частотный список нового типа - список частот, где словоформы и лексемы только с одной мотивацией объединены в одну общую структуру. Ранговая позиция модели определена как количество частот словоформ, которые ее составляют. Мотивированный частотный список связан с распределительным списком со ссылкой на все ситуации, где определенная словоформа появляется в определенном узле.

Systematic Arrangement of Terms in the Dictionary: Quantitative Approach to Linearization

S.D.Shelov,
Russia's Open University,
Department "Languages and Cultures", Moscow
Fax: 292 65 11 (box 3502 YEGRES)

Topical paper

AREA: Conceptual structure of terminological system

SUMMARY:

Is it possible that a linear succession of terms could be steadily based on their conceptual system? Could we imagine a term dictionary which is very easy to read as its items are arranged according to their meanings? The paper is supposed to discuss these problems.

There is one question which is always to be answered in compiling a term dictionary, no matter what a subject field, size, potential user and other parameters of the dictionary are. The question runs as follows: how dictionary's items should be arranged? Thus, according to A.J. Shajkevitch, this is one of the problems any lexicographer could run across in choosing a macrostructure of a term glossary (1, p. 25). He noted that a systematic arrangement of terms (i. e. arrangement of terms according to their meanings) is extremely rarely used (1, p. 29). This view was severely critisized by T.L.Kandelaki who stated that many vocabularies (collections of recommended terms, in particular) arrange terms systematically (2, p. 91). In a lot of these publications listed terms are really declared to be given in a systematic order, determined by classification of concepts as it is adopted in the corresponding field of knowledge. Yet the actual sense of these declarations (to say nothing of the exact sense) remains somewhat vague.

In fact, term lexicography has been discussing the problem of systematic arrangement of vocabular items. But there is no convincing answer to this problem either in terminology, or in lexicography. Moreover, in my opinion this problem has been hardly put forward in strict terms to give impetus to its solution (a recent article by D.F.Podpolny and E.F.Skorohodko (3), seems to be one of the few exceptions to the rule).

Though a linear order is a very well defined and a deeply investigated structure, obviously it does not meet a non-linear conceptual structure of any term system. Taking this into consideration, anyone wouldn't be surprised by the fact that up to now the systematic (i.e. semantic!) arrangement of terms is a purely intuitive procedure and the alphabetical order is the only strictly based way of arranging terms in a glossary. Meanwhile terminology has got rich experience of "conceptual linearization" of terms. Generalized in a theoretical framework and combined with some models of semantic structures, this experience could have put the question under consideration on mathematical basis and opened perspectives for its objective solution.

Two aspects of the problem seem to be worth mentioning. The first aspect (let me call it idcographic) is determined by categorization (classification) of all terminological units to be listed in a glossary, by fixing logical and conceptual relations between these categories and by choosing on these basis correct linear succession of these categories. For example, T.L.Kandelaki, speaking of the systematic arrangement of terms, follows this way: she picks out categories, blocks, sections of terms, determines interrelations between them and describes possible linearization of these categories, blocks, sections (4).

It should be noted that here we do not arrange terms themselves but rather large categories of terms. Besides, arguments in favour of the particular order of term categories lie in the plane of related subject field, theory of science, philosophy and even theory of culture. That's why some reasons for particular arrangement of lexical items in an ideographic dictionary are of greate importance for the topic and should be taken into consideration (5, 6). Ideographic approach is apparently necessary as we start to analyze a large and geterogeneous terminology corpus including terms of different sciences and disciplines (cf. terms of biology, philosophy, chemistry etc.) and we need its classification into smaller and more homogeneous groups of terms.

The second aspect (let us call it properly conceptual or semantical) is determined by linearization of proper terms within some fixed category of homogeneous terms according to logical-conceptual system of these terms. Here we do arrange relatively homogeneous terms according to their meanings. Analysis of the available glossaries in which term-ordering pretends to be "conceptual" ("semantic") leads to the exposure of three regular rules. These rules can be characterized as rules of semantic arrangement of terms and are formulated as follows:

1) if a term B is defined directly or inderectly through a term A, then the term B follows the term A, but not vice versa (the rule of definability);

2) if a term B is linguistically derived from a term A (in morphological or syntactical aspects of derivation), then, the first rule being carried out, the term B follows the term A, but not vice versa (the rule of language derivation);

3) if, in accordance with some semantic representation, for all or some pairs of terms we can determine the "concept (semantic) distance", then, the first and the second rule being carried out, the systematic arrangement of terms minimizes the total difference between concept (semantic) istance of terms as it appeares in a term list and that as it appeares in the semantic representation used (the rule of conceptual (semantic) distance).

The first two rules seem to be quite natural and can be inferred from the glossaries, mentioned by T.L.Kandelaki (in particular, the collectons of recomended terms ("Sborniki rekomenduemyx termiov"), published by the

Committee of the Scientific and Technical Terminology of the Academy of Sciences).

As far as the rule 3 is concerned, let me now start from the assertion that a linear order of terms naturally determines the distance between any two terms: that is, if a term A is listed as i-th and a term B is listed as j-th, the distance between A and B (D(A,B)) is equal to ABS(i-j). So for every given linear order of terms L we can calculate the total distance between any fixed set of terms and compare it with the total distance between the same set of terms in the semantic representation chosen. Afterwards the general idea of the rule 3 is theoretically transparent: in a term list terms should on the whole be separated, as accurate as possible, at the distance, fixed by the semantic representation. Suppose then we have a semantic graph G representing the corresponding terminology with the edges marked by the "distance figures" (that is of common practice nowadays), i.e. G(i,j)=D(i,j). Then the rule 3 is easily formulated in strict terms: among all the linear orderings of the terms we are to chose the one (or all of them if there are many) that minimizes the function

$$\sum_{i,j} |G(I,J)-ABS(I,J)|.$$

There is no doubt that resulting effect will greatly depend on the semantic representation used (and consequently of the distances between the terms represented). But still, supplied with the rules 1-3, we receive comparatively reliable criteria of what a semantic arrangement of terms could be.

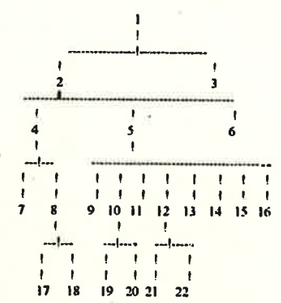
Let me now illustrate what has been said above by an example. Suppose we have the following list of 22 terms covering the subject field of technology "Pumps":

1. Pump, 2. Volume pump, 3. Dynamic pump, 4. Impeller pump, 5. Friction pump, 6. Electromagnetic pump,
7. Centrifugal pump, 8. Axial pump, 9. Scoop pump, 10. Peripheral pump, 11. Free peripheral pump, 12. Inclined Archimedian screw pump, 13. Disk pump, 14. Vibration pump, 15. Jet pump, 16. Inclined disk pump, 17. Fixed pitch-blade pump, 18. Adjustable pitch-blade pump, 19. Closed-peripheral pump, 20. Opened-peripheral pump, 21. Maze pump, 22. Worm pump.

Let these terms be defined only by gender-species definitions and, thereof, let these terms be semantically organized in a purely gender-species way as it is shown at fig.1.

Fig. 1.

Concept structure of the terms of the subject field "Pumps"



Note that only gender-species relations are taken into account here, so

D(pump, dynamic pump) = D(volume pump, friction pump) = D(friction pump, vibration pump) = 1,

while

D(pump, friction pump) = D(friction pump, maze pump) = D(volume pump, axial pump) = D(volume pump, centrifugal pump) = D (impellar pump, fixed pitch-blade pump) = 2;

D(volume pump, opened-peripheral pump) = D(volume pump, worm pump) = 3

and

D(dynamic pump, volume pump) = D(dynamic pump, electromagnetic pump) = D(friction pump, fixed pitch-blade pump) = + (INF)

What is then a semantic arrangement of these terms?

It may be proved that one solution is the following:

1. Pump, 2. Dynamic pump, 3.Volume pump, 4. Electromagnetic pump, 5. Impeller pump, 6. Centrifugal pump, 7. Axial pump, 8. Fixed pitch-blade pump, 9. Adjustable pitchblade pump, 10. Friction pump, 11. Scoop pump, 12. Free peripheral pump, 13. Disk pump, 14. Vibration pump, 15. Jet pump, 16. Inclined disk pump, 17.Peripheral pump, 18. Closed-peripheral pump, 19. Opened-peripheral pump, 20. Inclined Archimedian screw pump, 21. Maze pump, 22. Worm pump.

The structural approach proposed above can be easily formalized in many ways mainly because of different possible semantic representations and explications of what "a conceptual (semantic) distance" between two terms is (sometimes even within one and the same semantic formalism). The optimization procedure is attributed to the integer programming of none-standard type and far from being trivial. It is supposed however that the general idea outlined above would present wide perspectives for semantic arrangement of terms in a dictionary on a steady ground of calculations. In some simpliest cases we can get clear and acceptable solutions just today.

References

1. Shajkevich A.J. Problemy terminologicheskoj leksikografii /Problems of Term Lexicography/. - Moscow: VCP, 1983. - 67 p. (Vsesojuznyj tsentr perevodov (VCP). Perevod nauchnoj i tehnicheskoj literatury. Ser. 1. Teorija i praktika perevoda. Obzor. - N 8. /All-Union Translation Center (VCP). Translation of Scientific and Technical Literature. Ser.1. Theory and Practice of Translation. Review. - N 8).

2. Kandelaki T.L. Slovari sistematicheskogo tipa i "Slovar slavjanskoj lingvisticheskoj terminologii" /Systematic Dictionaries and "Dictionary of Slav Linguistic Terminology" // Slavjanskaja lingvisticheskaja terminologija / Slav Linguistic Terminology /. - Kiev, 1984. - P. 88 - 96.

3. Podpolny D.F., Skorohodko E.F. Ob opredelenii znachimosti slova /On determining word value //Nauchno-tehnicheskaja informatsija /Scientific and Technical Information/. Ser. 2. - 1989. - N 2. - P. 9 - 15.

4. Kandelaki T.L. Rol terminologicheskih slovarej sistematicheskogo tipa v processe izuchenija osnov nauk studentami-inostrantsami /Term Dictionaries of Systematic Type in Studing Bases of Sciences by Foreign Students //Problemy uchebnoj leksikografii /Problems of Educational Lexicography/. - M., 1977. - P. 150 - 157.

5. Karaulov J.N. Obschaja i russkaja leksikografija /General and Russian Lexicography/. - Moscow: Nauka, 1976. - 355 p.

6. Morkovkin V.V. Ideograficheskie slovari /Ideographic Dictionaries/. - Moscow, 1970.

7. Shelov S.D. O principah numeratsii terminov v terminologicheskih standartah i sbornikah rekomenduemyh terminov /Sratetegies for Arrangement of Terms in a Normative Terminology Dictionary //Nauchno-tehnicheskaja terminologija / Scientific and Technical Terminology /. - 1986. - N 3. - P. 5-9.

8. Shelov S.D. O logico-smyslovom raspolozhenii terminov /na materiale terminologicheskih standartov i sbornikov rekomenduemyh terminov /On Semantic Arrangement of Terms (for Terminology Standards and Collections of Recommended Terms) // Nauchno-tehnicheskaja terminologija /Scientific and Technical Terminology. - 1988. - N 6. - P. 6 - 9.

Систематическая организация в словаре: квантитативный подход к линеаризации

Шелов С.Д.

Резюме:
Возможно, что линейная последовательность терминов могла быть основана на их понятийной системе. Можем ли мы представить словарь терминов, который очень легко читать, т.к. его статьи (параграфы) расположены согласно их значениям ? Настоящая статья предполагает обсуждение этих вопросов.

Automatic Typological Analysis of Semitic Morphology

Arthur V. Stepanov

Institute of Oriental Studies, Russian Academy of Sciences,
Rozhdestvenka 12, Moscow 103777, Russia

Mailing address: ul. marshala Vasilevskogo 3/1–17, Moscow 123098, Russia;
phone: (+7 095) 196-63-62;
e–mail: arth@aestep.msk.su

The Greenbergian quantitative approach to the morphological typology developed through automation can be effectively used for a fine analysis of a natural language morphology. In this paper the automatic analyzer is realized for the case of Semitic morphology and its use is exemplified with some of the Semitic languages, namely, Literary Arabic and Maltese.

Topical paper

Specification: quantitative linguistics

Automatic Typological Analysis of Semitic Morphology

Arthur V. Stepanov

The traditional morphological typology describes languages on the basis of comparing their morphological features. To obtain the detailed and precise typological characteristics one must strictly consider the degree of representation (relevancy) of one or another morphological feature in a language. The well–known quantitative typological approach of J.H. Greenberg (1960) enables to express such a degree through calculation of the appropriate numerical indices, thus adequately accounting for the presence of opposite morphological features (synthesis and analysis; agglutination and fusion etc.). However, there is a number of theoretical linguistic problems, a solution of which requires more fine analysis of a language structure with regard of dynamical changes taking place inside it. These problems are, for instance, the change from synthetical to analytical structure of morphology in some languages, stability and variation of the morphological features, fine morphological differences in the related languages, the general and specific tendencies in the development of natural language grammars etc. Here we will attempt to demonstrate by the example of some Semitic languages, how the Greenberg's approach established on the statistical basis and developed through automation could be used, beside a formal typological classification of languages, for the fine analysis of the morphological structure of a language with respect to all language variety.

1 Problem and Background

To adequately describe a morphological feature through its degree of representation in sample texts we have to successively proceed from the statistical nature of the appropriate index within the framework of the Greenberg's definition. That means we should get rid of any subjective factors while operating with texts and accept an assumption about the random and mass character of text selection.[1]

Following this approach we assume that values of any index computed over the different text units show natural dispersion which bears both the random and systematic character and depends, along with index average, on the language or dialect type, time of being written, stylistic register of a text and many other reasons. Since those reasons are objective, we infer that a morphological feature is described not only by the appropriate index average but also by the index distribution function as a whole. The large–scale analysis of texts in the language under investigation is to show whether this distribution function is normal, and, if no, then what systematic factor disturbs the random dispersion for that feature and makes differentiation inside the morphological structure.

To properly process great numbers of texts an automation is needed. The computer program especially elaborated for this purpose and compiled in PROLOG provides with the automatic quantitative typological analysis of natural language texts on the morphological level. This permits to obtain the Greenbergian index data arrays to be subsequently processed by the mathematical statistics methods. At present the program using an algorythm formalizing the Semitic morphological structure (described by the approach of Yushmanov 1961) performs the analysis of texts written, primarily, in Literary Arabic and some Arabic dialects, including Maltese.

Here we are dealing with only five typological indices defined after Greenberg: index of synthesis, prefixation index, suffixation index, gross inflectional index, derivational index.

[1] The existing quantitative typological data, including those for Afro–Asiatic languages are based on analysis of a relatively small number of texts (Krupa 1965 and Khrakovsky 1982).

2 Applications and Results

I. In the case of Literary Arabic, we chose 90 sample texts of about 700 words each. The number of texts was determined by the required accuracy for computing every index average (no less than 95%). For the text selection to be random and objective the texts have been chosen from the different stylistic registers (scientific, social, belles–lettres), sources (books, newspapers, magazines), though all related to the modern period.

The quantitative statistical data obtained for the above indices are illustrated in Fig.1.

We can see that the synthesis and derivational indices (histograms in Fig.1(a) and (b)) are not distributed normally, while the prefixation, suffixation and gross inflectional indices show the normal (random) distribution (Fig 1(c), (d) and (e)). That means in the context of the features of syn-

thesis and derivation (word–formation) the structure of morphology can be differentiated by a more specific criterion.

Our study shows in the case of Literary Arabic a stylistic differentiation should be considered for such a criterion. Indeed, the synthesis and derivational index data for the texts separated by the different stylistic registers are distributed normally. The left dotted curves in Fig.1(a) and (b) exhibit the appropriate index distributions over the texts belonged to the belles–lettres, while the right curves exhibit the distributions over the texts belonged to both social and scientific stylistic registers. The synthesis and derivation (word–formation) in scientific texts are higher than in belles–lettres, which displays the more intensive semantic and grammatical load of a word. We thus can compute precisely how many times one stylistic register is separated from the other for the given morphological feature.

In this respect, it is important for typology to establish what indices, i.e. properly what morphological features are influenced by the stylistic differentiation essentially and which are not. In our example of Literary Arabic, obviously, this differentiation affects the synthesis and derivation (word–formation) and does not affect the inflection, prefixation and suffixation. In other languages the differentiation can influence morphology in another way. Besides, for any languages the nature of differentiation is likely to differ from the stylistic one. Nevertheless, it is apparent that the morphological features whose degree of representation does not depend on stylistic (or any other kind of) differentiation are the most stable and, therefore, the most valuable ones from the typological point of view.

II. The process of revealing differences in the morphologies of the related languages by way of the fine statistical analysis can be demonstrated by the example of Literary Arabic and Maltese. In our study we involved about 30 texts of about 700 words each from Literary Arabic and the same number of texts from Maltese. For a text analysis to be adequate enough the texts have been chosen from one stylistic register (social). The accepted accuracy of computation was no less than 95%.

The results of computing the Greenbergian indices for Maltese in comparison with Literary Arabic are shown in Table 1.

Referring to Table 1, we can see that Maltese morphology is clearly separated from Arabic in such extent that the differences in index averages are in great excess of the natural dispersion for each of them. The evident

Table 1. Greenbergian typological indices for Literary Arabic and Maltese.

Index	Literary Arabic		Maltese	
	Average	STD deviation	Average	STD deviation
Synthesis	2.58 ± 0.03	0.1	1.72 ± 0.02	0.05
Derivation	0.60 ± 0.02	0.09	0.31 ± 0.01	0.05
Inflection	0.98 ± 0.02	0.06	0.41 ± 0.02	0.04
Prefixation	0.50 ± 0.02	0.05	0.29 ± 0.01	0.04
Suffixation	0.96 ± 0.03	0.09	0.30 ± 0.01	0.05

reason for these differences consists in the Indo European impact on the Maltese grammar (Acquilina 1973), which proves to be significant enough to appreciably change originally Semitic based morphological structure of Maltese.

The described method of the automatic typological analysis with employing the Greenbergian quantitative approach can be prospectively applied in other aspects, where revealing the fine structure of morphology is required:

1. In our case the analysis of the Arabic morphology falls into the differentiated analysis of morphological substructures governed by stylistic registers. For other languages the main criterion for differentiation is likely to be another one, for example, temporal (historical) or geographical. We can provide a precise mathematical description for the differences in morphologies by one or another criterion.

2. In this regard, depending on the criterion for the differentiation we would involve the automatic identification of the stylistic register, area or historical period related to the texts being analyzed.

3. If that criterion is the temporal (historical) one, we can, on the base of reliable index data, follow the dynamics of language development and to derive a "typological function" of the language on time.

4. If we consider typological characteristics in the context of grammaticalization of the human thinking verbalization process (Kibrik 1990), the Greenbergian indices computed by the described method could also be involved there.

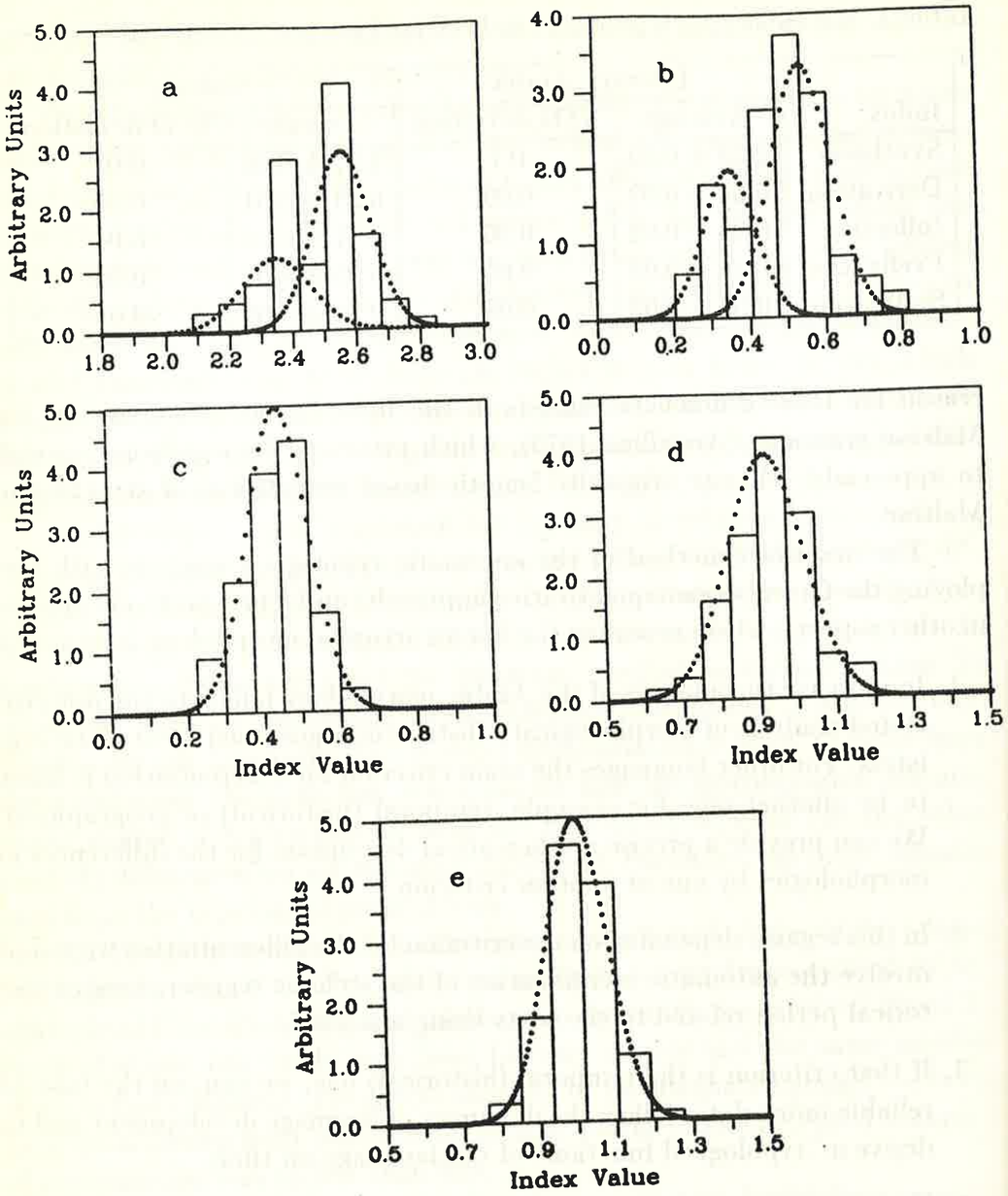


Fig. 1. Distribution density of the Greenbergian indices for Literary Arabic: (a) index of synthesis; (b) derivational index; (c) prefixation index; (d) suffixation index; (e) gross inflectional index.

References

Acquilina, J. (1973). *The Structure of Maltese: A Study in Mixed Grammar and Vocabulary.*, Msida: the Royal University of Malta.

Greenberg, J.H. (1960). "A quantitative approach to the morphological typology of languages." *International Journal of American Linguistics*, 26(3): 178–194.

Khrakovsky, V.S. (1982). "The Literary Arabic language". In *Quantitative Typology of the Afro-Asiatic Languages)*, edited by V. Kasevich and S. Yahontov, 292–305. St. Petersburg: University press. (In Russian)

Kibrik, A.E. (1990). Artificial intelligence and linguistic typology. In *Proceedings, All-USSR Conference on Linguistic Typology*, 22–25. (In Russian)

Krupa, V. (1965). "On quantification of typology." *Linguistics*, 12: 31–36.

Yushmanov, N.V. (1961). *The Structure of the Arabic Language* (transl. Moshe Perlmann.), Washington D.C.: Center for Applied Linguistics.

Disambiguation by Association: Two Algorithms and their Assessment

Richard F.E. Sutcliffe .
Bronwyn E.A. Slater

Department of Computer Science
and Information Systems
University of Limerick
Limerick, Ireland
+353 61 333644 Ext 5006
+353 61 330876 (FAX)
sutcliffer@ul.ie
slaterb@ul.ie

Word sense disambiguation is vital to accurate text analysis. We have replicated two well known methods due to Lesk (1986) and Ide and Veronis (1990) and made some interesting discoveries. Firstly, disambiguation is a much more complex and subtle problem than is generally assumed. Secondly, the performance of a particular algorithm is heavily dependent on the way in which it is measured. Thirdly, the Ide and Veronis method does not perform significantly better than the Lesk technique. We conclude by proposing a number of metrication factors which could lead to more representative disambiguation results.

Topical Paper

Topic Areas: computational lexicology, word sense disambiguation, neural networks, objective performance metrics in NLP systems

1 Introduction

The ability to disambiguate the senses of each word in a text is vital if its meaning it to be determined accurately. In *Animal Farm* by George Orwell for example, 87% of words have more than one sense[1]. With the advent of machine readable dictionaries (MRDs) various ingenious methods have been proposed to disambiguate words automatically. Two well-known techniques are the Lesk method (Lesk, 1986) and the Cottrell-Veronis-Ide (CVI) method (Veronis and Ide, 1990). [2] Both of these exploit the idea that the correct senses of a pair of words in a sentence will be semantically related and that this can be detected using their definitions. We have replicated these methods using the same dictionary and test corpus for each and have made a number of interesting discoveries. We report on these below.

2 The Disambiguation Algorithms

2.1 The Disambiguation Task

In a word sense disambiguation task the objective is to assign to each word in a text an appropriate sense chosen from a particular MRD. Thus to disambiguate "pen paper" relative to the Merriam-Webster Compact Electronic Dictionary we choose sense three of 'pen' ("tool for writing with ink"), and sense one of 'paper' ("pliable substance used to write or print on, to wrap things in, or to cover walls").

2.2 The Lesk Method

The Lesk disambiguation method involves the use of frequency counts in computing the preferred sense of each word in the input phrase or sentence. Firstly, all the sense definitions of each word in the input are looked up in the dictionary. Analysis then proceeds by discarding each word in a sense definition which does not occur in any other definition. Each remaining word in a definition is converted to its root inflection. A count is made of the number of times it occurs in other definitions and that count is then associated with the word wherever it occurs. A score is then determined for each sense definition by computing the product of the word scores within it. Finally, each word is disambiguated by choosing the sense which has the highest score.

2.3 The Cottrell-Ide-Veronis Method

Cottrell-Ide-Veronis disambiguation is similar in spirit to the Lesk method but uses a spreading activation network with two-way arcs. There are two types of node in the network, *word* nodes and *sense* nodes. The network is created by first allocating one word node to each content word in the input (function words are eliminated). Thus for "pen paper" one word node is allocated for 'pen' and another for 'paper'. Each word node is connected by excitatory arcs to sense nodes, one for each semantic sense of the word as defined in the dictionary. Thus we might have 'pen' connected to pen1, pen2, pen3 and pen4, with 'paper' connected to paper1-paper5. Each set of sense nodes for a word is strongly interconnected by inhibitory arcs to form a winner-take-all network. Each sense node is then connected to one word node for each word occurring in that sense definition, converted to its root inflection. Thus pen3 would be

[1] According to our preliminary findings on a corpus of 100 sentences

[2] Two very interesting corpus based approaches to disambiguation are those of Schuetze (1993) and Gale, Church and Yarowsky (1993). The authors report very good results for these methods. However they are not readily applicable to a task whose objective is to point to the correct sense in a dictionary.

Table One: Corpus Ambiguity

Category	Word Counts vs. Degree of Ambiguity									
	1	2	3	4	5	6	7	8	9	10
Nouns	85	91	106	49	63	20	15	13	22	5
Verbs	30	56	50	38	41	20	30	17	7	9
Adjectives	17	15	48	28	17	9	11	6	7	3
Adverbs	7	1	12	4	3	15	11	7	0	0

connected to 'tool', 'write' and 'ink'. These links are excitatory. There is only one instance of each word node in the network. Thus if a word occurs in more than one definition, several sense nodes will be connected to it. Because these nodes join different parts of the network, the system can capture the semantic links between senses of different words in the input.

The activation functions used in the network are very standard. The activation at time $t+1$, $a_i(t+1)$ is defined as follows:

$$a_i(t+1) = a_i(t) + s_i - \delta \quad (1)$$

The squashed net input s_i is defined by

$$s_i = n_i(1 - a_i) \quad when\ n_i > 0$$
$$s_i = n_i a_i \quad when\ n_i < 0 \quad (2)$$

where the net input to node i, n_i is

$$n_i = \sum_j w_{ji} a_j \quad (3)$$

Decay δ is given by

$$\delta = D_1(a_i - D_2) \quad (4)$$

where D_1 and D_2 are constants.

After the network has been created, the cycling phase begins. The activation of the input word nodes is set to 0.2 and the network is run until a situation of stability has been reached. Words which occur in sense definitions of several different input words will tend to become more active because they receive input from more than one part of the network. As a result they will tend to reinforce the sense nodes to which they are connected, thus pushing down competing senses. Disambiguation is accomplished by choosing from each winner-take-all network the sense which has the highest activation.

The network described above is of height one (CVI-1). A CVI-2 network can be created by taking each word node which occurs at the bottom of the network and creating further nodes for it. Firstly, we create a sense node for each sense of that word in the dictionary. Secondly, we add word nodes under each sense node corresponding to the words which occur in the definition of that sense, just as before. In general, a CVI network of any height can be created by repeating this process.

3 Disambiguation Trials

3.1 The Tests

First of all, a corpus of 100 sentences was created from *Animal Farm* by George Orwell. Function words were eliminated. Each word was then disambiguated manually by two human

Table Two: Disambiguation Performance Results

	Lesk	CVI-1	CVI-2	CVI-3
Total sentences	100	100	100	27
Total words	2647	2647	2647	679
Total content words	1094	1094	1094	289
Total ambiguous words	954	954	954	250
	(87%)	(87%)	(87%)	(86%)
Total unambiguous words	140	140	140	39
	(13%)	(13%)	(13%)	(14%)
Total correct	422	660	651	168
(includes unambiguous words)	(39%)	(60%)	(59.5%)	(58%)
Total ambiguous correct	282	520	511	129
	30%	55%	54%	52%
Total ambiguous isolated	412	412	0	0
	(43%)	(43%)		
Total ambiguous non-isolated	542	542	954	250
	(57%)	(57%)	(100%)	(100%)
Total ambiguous	282	276	511	129
non-isolated correct	(52%)	(50%)	(54%)	(52%)

subjects. During the disambiguation session, a subject was presented with a complete sentence on the screen together with the appropriate definitions from the Merriam-Webster Compact Electronic Dictionary. They then selected zero, one or more senses for each word which they considered appropriate for its use in that context. The results of each session were saved in a file. The 'correct' set of senses for each word in a given sentence was then created by taking the intersection of the sets created for it by the pair of subjects. The result was a set of 1094 disambiguated content words which were then used for testing the algorithms. This compares favourably with the 138 word pairs used in the Veronis and Ide study. Table One provides some information about ambiguity in the corpus. Specifically it provides frequency counts of 1-way to 10-way ambiguous words broken down over syntactic category.

The corpus was then used to test four algorithms, Lesk, CVI-1, CVI-2, and CVI-3. The results of the tests are summarised in Table Two. The terms used in column one can be explained as follows. **Content words** are defined to be those of category noun, verb, adjective or adverb. Function words are thus excluded. An **ambiguous word** has more than one sense in the dictionary while an **unambiguous word** has only one sense. **Total correct** is a count of the ambiguous words which were disambiguated correctly *plus* a count of all the unambiguous words. **Total ambiguous correct** is the real test of the algorithms. It is the number of ambiguous words which could be disambiguated correctly. **Isolated words** are those whose definitions share no words with other definitions in the sentence being disambiguated. By definition, such words can not possibly be disambiguated by either an Ide or a Lesk method. **Total ambiguous isolated** is a count of the ambiguous words which are isolated. **Total ambiguous non-isolated** is a count of the ambiguous words which are in principle disambiguatable by the methods. Finally, **total ambiguous non-isolated correct** is a count of the disambiguatable words which were correctly chosen by the methods. This is the true measure of performance of the algorithms.

Our principal finding is that all the methods give broadly comparable results of 50-54 percent ambiguous non-isolated correct. CVI-2 gives the best result at 54 percent. The difference between the Lesk ambiguous correct and non-isolated ambiguous correct figures is because

Lesk can not disambiguate isolated words at all. The Ide methods disambiguate such words at chance, giving a superficially better performance.

Ide and Veronis report a higher figure of 72 percent in their study (Ide and Veronis, 1990). However their work was conducted on word pairs rather than complete sentences which is an easier task. Also it is important to note that we included all syntactic categories in our experiments - noun, verb, adjective and adverb - and that we only considered an algorithm to have correctly disambiguated a particular word if it chose the right sense *and* the right syntactic category. This is a severe test of an algorithm.

It is interesting to note that in this study the CVI-3 networks did not perform better than CVI-2 networks. In addition CVI-3 networks were considerably larger in size, comprising around 3000-4000 nodes and 10,000 bidirectional arcs. This suggests that the "interesting" words occur in the more immediate dictionary definitions rather than in those at a deeper level, implying that CVI-3 networks may not be worth the extra space and time requirements which they incur.

3.2 Conclusions

The main conclusion of this study is that the methods all perform at a comparable level and that the spreading activation technique is not superior to the Lesk (word intersection) method.

However, our work indicates that the results of a disambiguation experiment have to be considered in the light of the way in which the experiment has been conducted. The following have been shown to be important factors:

- **1-way ambiguous words.** Whether these are included in the statistics makes a large difference to the result.
- **Intrinsically undisambiguatable words** (e.g. isolated words in our studies). The fact that these can be detected at run time with the algorithms discussed here is very useful from the perspective of later semantic processing of a text. However they should be excluded from the statistics.
- **The number of words disambiguated at a time.** The use of whole sentences makes the disambiguation task more difficult but it is seems a likely way in which an algorithm would be used in a text processing application.

Factors not investigated in this study include:

- **The effect of the dictionary used on results.** We used the same dictionary for all trials, namely the CED. It is possible however that other dictionaries could give a higher level of performance overall or that they particularly suit a given algorithm.
- **The domain of the corpus.** The particular application domain in which the disambiguation is to be used may well affect results. In addition, higher levels of performance can undoubtedly be obtained in a restricted domain, for example by exploiting domain-specific word-sense frequency data. For example in a computer manual 'file' almost certainly means a computer file.
- **Criteria for the selection of test sentences.** Undoubtedly the size of the corpus and its composition in terms of sentence length, proportion of function words and so on will affect results.

- **The proportion of words *in principle* disambiguatable by a semantic correlation method.** The methods described here, although powerful and simple, can clearly not disambiguate every ambiguous word even in principle. For example in 'Pick up a file' it is impossible to say whether the file is a document file or a mechanic's file, because the other words in the sentence give no clues via any semantic correlation. We need to know what proportion of words in a corpus are in this category. For example, it might be the case that most of the words which the algorithms disambiguated incorrectly are in fact of this kind. This would imply that the semantic correlation methods were in fact working at a very high level of performance, thus necessitating the investigation of new methods rather than the optimisation of existing ones.
- **The effect of syntactic category ambiguity.** In this study we forced the algorithms to choose the correct syntactic category as well as the correct sense. If a tagger was used to select category before performing disambiguation this would improve the results.

Clearly there are many interesting avenues for this work and we are currently engaged in researching some of the above issues.

4 References

Cottrell, G.W. (1985). *A Connectionist Approach to Word Sense Disambiguation.* Doctoral Dissertation, Department of Computer Science, University of Rochester.

Gale, W. A., Church, K. W., & Yarowsky, D. (1993). A Method for Disambiguating Word Senses in a Large Corpus. *Computers and the Humanities*, **26**(5-6), 415-439.

Ide, N. M, & Veronis, J. (1990). Very Large Neural Networks for Word Sense Disambiguation. *Proceedings of the European Conference on Artificial Intelligence, ECAI'90, Stockholm, August 1990.*

Lesk, M. (1986). Automated Word Sense Disambiguation using Machine-Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. *Proceedings of the 1986 SIGDOC Conference.*

Schuetze, H. (1993). Word Space. In S. J. Hanson, J. D. Cowan and C. L. Giles (Eds.), *Advances in Neural Information Processing Systems 5.* San Mateo CA: Morgan Kaufmann.

Figure 1. Synset Hierarchy for the word 'terrier' derived from Princeton Wordnet.

```
1 sense of terrier

Sense 1
terrier --
   (any of several usu. small short-bodied breeds originally trained
    to hunt animals living underground)

 => hunting dog --
      (a dog used in hunting game)

  => domestic dog, pooch, Canis familiaris --
       (domesticated mammal prob. descended from the common wolf; occurs in
        many breeds)

   => dog

    => canine, canid --
         (any of various fissiped mammals with nonretractile claws and
          typically long muzzles)

     => carnivore --
          (terrestrial or aquatic flesh-eating mammal; terrestrial carnivores
           have four or five clawed digits on each limb)

      => placental mammal, eutherian, eutherian mammal

       => mammal --
            (any warm-blooded vertebrate that nourish their young with milk
             and having the skin more or less covered with hair; young are
             born alive except for the small subclass of monotremes)

 => vertebrate, craniate --
       (animals having a bony or cartilagenous skeleton with a
        segmented spinal column and a large brain enclosed in a skull
        or cranium)

  => chordate

   => animal, animate being, beast, brute, creature, fauna --
        (a living organism characterized by voluntary movement)

    => life form, organism, being, living thing --
         (any living entity)

     => entity --
          (something having concrete existence; living or nonliving)
```



```
repn( terrier, '(any of several usu. small short-bodied breeds originally
trained to hunt animals living underground)', [[any, 0.19], [several, 0.19],
[small, 0.19], [breed,0.19], [originally, 0.19], [trained,0.19], [hunt, 0.19],
[animal,1.9], [living, 0.19], [underground, 0.19], [a, 0.17], [dog, 0.17],
[used, 0.17], [in, 0.17], [hunting, 0.17], [game, 0.17], [domesticated, 0.15],
[mammal, 0.15], [descend, 0.15], [common, 0.15], [wolf, 0.15], [occur, 0.15],
[many, 0.15], [various, 0.11], [fissiped, 0.11], [with, 0.11], [nonretractile,
0.11], [claw, 0.11], [typically, 0.11], [long, 0.11], [muzzle, 0.11],
[terrestrial,0.096], [aquatic, 0.096], ['flesh-eating', 0.096], [carnivore,
0.096], [have, 0.096], [four, 0.096], [five, 0.096], [clawed, 0.096], [digit,
0.096], [on, 0.096], [each, 0.096], [limb, 0.096], ['warm-blooded', 0.057],
[vertebrate, 0.057], [nourish, 0.057], [young, 0.057], [milk, 0.057], [skin,
0.057], [more, 0.057], [less, 0.057], [covered, 0.057], [hair, 0.057], [are,
0.057], [born, 0.057], [alive, 0.057], [except, 0.057], [monotreme, 0.057],
[bony, 0.038], [skeleton, 0.038], [segment, 0.038], [spinal, 0.038], [column,
0.038], [large, 0.038], [brain, 0.038], [enclosed, 0.038], [skull, 0.038],
[cranium, 0.038], [organism, 0.01], [characterized, 0.01], [voluntary, 0.01],
[movement, 0.01], [entity, 0.01], [concrete, 0.01], [existence, 0.01],
[nonliving, 0.01]] ).
```

Figure 2. The semantic representation for 'terrier' produced by the algorithm.

remaining content words are converted to their root inflection. All such words are considered to be features of the word-sense, and are given a centrality of 1.0. We then chain upwards using a hypernymic link (if any)[1]. At the next level up, features are extracted from the hypernym's gloss, using a centrality of 0.9. The process is repeated, reducing the centrality by 0.1 at each level, until either the top of the hierarchy is reached or the centrality falls to zero. Finally, the representation, consisting of a set of feature-centrality pairs, is normalised.

3 Results

We have implemented the above algorithm, generated some noun representations and carried out initial tests on the results. Figure 1. Shows the synset hierarchy for the word 'terrier' as defined in WordNet. Figure 2. shows the distributed semantic representation which the algorithm produces for that word. One way in which the lexicon can be evaluated is to compute word-word similarity measures for a given set of word pairs and then to analyse the results for plausibility. In an initial experiment we chose five categories of concept, cars, dogs, flowers, trees and people. Four words were chosen within each category to use in our tests (Table 1.). Three pairs of categories were then chosen, cars-dogs, flowers-trees and people-dogs. Each category pair contains four words from the first category and four from the second, eight words in total. An eight-by-eight matrix of word-word similarity figures was then drawn up for each category pair (Tables 3-5).

There are several points to note about these. Firstly, in Table 3 the match of one car word with another is high, ranging between 0.58 and 1.0 with an average of 0.8. This shows that the lexicon has captured the similarity between the car concepts. Secondly, the match of one dog word with another is also high, ranging between 0.63 and 1.0 with an average of

[1] At present we only choose the first such link if there are several.

Table 1. Twenty words in five categories

cars	dogs	flowers	trees	people
chariot	pug	pansy	larch	bruiser
motorbike	terrier	daffodil	pine	patriarch
jeep	lapdog	tulip	oak	siren
moped	chihuahua	rose	sycamore	rake

Table 2. Lexical Representation Summary

No of words	20
Total number of features	249
Average number of features	39
Minimum	17
Maximum	76

0.76, for the same reason. Thirdly, the match of a car word with a dog word is low, ranging between 0.05 and 0.17 with an average of 0.1. This is because cars and dogs are not closely linked semantically. Table 4 shows results for the flowers-trees matrix. Flowers and trees are much more closely related semantically than cars and dogs, and this is reflected in the results. Flower words match with tree words in a range of 0.30 to 0.67 with an average of 0.4, much higher than for cars and dogs. The match of flowers with flowers or trees with trees continues to be high. Finally, Table 5 shows the people-dogs matrix. Note here that the match of people with themselves is lower than that of dogs with themselves (average 0.63 rather than average 0.76.) This is because the people words are in fact a rather disparate set. Note in particular that 'bruiser' against 'rake' is the best match while 'bruiser' against 'patriarch' is the worst. This matches one's intuitions about these concepts: patriarchs are "good" while 'bruisers' and 'rakes' are not.

4 Conclusions

We have presented one algorithm by which a large lexicon of distributed semantic representations for nouns can be generated automatically from the Princeton WordNet. We have also shown some initial tests which support the hypothesis that these representations are capturing the meanings of the words. While simple vectors can not capture every aspect of meaning, the wide coverage of a lexicon of this kind renders it a promising candidate for applications such as information retrieval where large quantities of unrestricted text need to be analysed with reasonable accuracy. We are at present refining our algorithms and investigating other strategies for measuring the performance of a lexicon objectively.

Table 3. Cars vs. Dogs

	chariot	motorbike	jeep	moped	pug	terrier	lapdog	chihuahua
chariot	1.00	0.74	0.58	0.73	0.13	0.17	0.14	0.09
motorbike	0.74	1.00	0.69	1.00	0.11	0.11	0.11	0.06
jeep	0.58	0.69	1.00	0.68	0.08	0.09	0.09	0.05
moped	0.73	1.00	0.68	1.00	0.10	0.10	0.11	0.05
pug	0.13	0.11	0.08	0.10	1.00	0.68	0.65	0.69
terrier	0.17	0.11	0.09	0.10	0.68	1.00	0.63	0.72
lapdog	0.14	0.11	0.09	0.11	0.65	0.63	1.00	0.67
chihuahua	0.09	0.06	0.05	0.05	0.69	0.72	0.67	1.00

Table 4. Flowers vs. Trees

	pansy	daffodil	tulip	rose	larch	pine	oak	sycamore
pansy	1.00	0.32	0.36	0.49	0.37	0.32	0.37	0.28
daffodil	0.32	1.00	0.70	0.37	0.38	0.33	0.37	0.39
tulip	0.36	0.70	1.00	0.41	0.39	0.33	0.37	0.30
rose	0.49	0.37	0.41	1.00	0.56	0.58	0.67	0.44
larch	0.37	0.38	0.39	0.56	1.00	0.83	0.74	0.64
pine	0.32	0.33	0.33	0.58	0.83	1.00	0.83	0.62
oak	0.37	0.37	0.37	0.67	0.74	0.83	1.00	0.60
sycamore	0.28	0.39	0.30	0.44	0.64	0.62	0.60	1.00

Table 5: People vs. Dogs

	bruiser	patriarch	siren	rake	pug	terrier	lapdog	chihuahua
bruiser	1.00	0.40	0.52	0.63	0.12	0.15	0.13	0.08
patriarch	0.40	1.00	0.40	0.55	0.15	0.18	0.16	0.17
siren	0.52	0.40	1.00	0.50	0.14	0.17	0.14	0.09
rake	0.63	0.55	0.50	1.00	0.12	0.15	0.13	0.08
pug	0.12	0.15	0.14	0.12	1.00	0.68	0.65	0.69
terrier	0.15	0.18	0.17	0.15	0.68	1.00	0.63	0.72
lapdog	0.13	0.16	0.14	0.13	0.65	0.63	1.00	0.67
chihuahua	0.08	0.17	0.09	0.08	0.69	0.72	0.67	1.00

Table 6 Word-Word Summary

Cars-Cars average	0.80
Cars-Dogs average	0.10
Dogs-Dogs average	0.76
Flowers-Flowers average	0.58
Trees-Flowers average	0.40
Trees-Trees average	0.78
People-People average	0.63
Dogs-Dogs average	0.76
People-Dogs average	0.14

words first(section 4.1). This distance shows us whether a certain word is resemble to the other words or not. We can get the distance from d-bigram, too. According to this distance, we get a group of words which elements are resemble to each other. We call this group a cluster of words. Our experimental result shows that the cluster comes to be practically a grammatical category. However, we do not use any grammatical information to cluster, so we dare to call this information just 'cluster'(section 4.2). Once we get clusters, it is not so difficult to smooth d-bigram data(section 4.3). All we have to do is that we calculate a new information of each elements in a cluster using all elements in the cluster.

Our system takes natural language sentence as its input, English, Japanese or Chinese. Both Japanese and Chinese are agglutinative, and we need the morphological analysis step. The method of the analysis is as complicated and important as the method of translation. An English sentence has several words and those words are separated with a space. It is easy to see how you can divide an English sentence into words. However, for example a Japanese sentence needs parsing if you want to pick up the words in the sentence. We divide agglutinative-language sentences into words(morphemes) without using any grammatical information. Instead, this system uses the statistic information between words(morphemes) to select best ways of segmenting sentences in agglutinative languages.

Lexical analysis is a simple procedure. After the morphological analysis, a right words which make up the sentence can be taken. We translate the words to the corresponded words in the target language. by using a word-dictionary. However, we often come to see the difference between languages at this point. For example a preposition in English is hard to translate into Japanese. Some should be translated in several ways, and some should not be translated. The same thing happens in English-Chinese translation too. This is a problem need to be solved, so we take some heuristic approach in this paper(section 5.2).

After we get words translated into target language, we construct a sentence according to a statistical information of that target language. The statistical information is characterized by some factors of the language. So we can generate the sentence which fits to the language. Brown used a trigram to generate the sentence in his paper[3]. Our method(section 5.3) based on d-bigram has two major efficiency as compared with a method based on trigram. First, a trigram is based on a sequence of just three words, so, words at a long distance, more than four words away, have no effect. Our method is based on d-bigram, so the information between words at long distance gives us better effect on generating a sentence. Secondly, when we want to generate a sentence which consists of three words, we have to have just the same word sequence in the trigram. In other words, when we do not have a certain sequence of three words in a corpus, we can not create the sequence at all. A d-bigram is more generalized information than trigram, so we do not need the exact sequence in a corpus. When word-B appears next to word-A in some place of the corpus and word-C appears next to word-B in other place of the corpus, we can create a sequence – word-A, word-B, word-C –.

3 Statistical Model

There are a lot of constraint on natural languages. For example, we can use 'wine', 'juice' and 'beer' after 'drink', however 'stone' is not suitable. In NLP, it is very hard to write these phenomena as a rule. We use statistical model to express these phenomena implicitly. In other words, we get these phenomena as statistical information because they exists normally in a corpus.

In statistical model we consider a probability of a word w_{n+1} which appears next to sequence $w_1, w_2, \ldots, w_n$. When n is three, it is called trigram model. Trigram showed a good result in automatical speech recognition, and some research for NLP based on trigram have been done[3]. Thus trigram is considerably good model for language, however, it is very hard to get enough trigram data for all sequences of certain three words.

3.1 D-bigram

We introduce our new statistical model named d-bigram. This is a kind of n-gram information, however we add 'distance parameter' to n-gram as new feature. D-bigram is fundamentally a bigram information. Normal bigram information takes the probability that a word w_1 is followed by a word w_2. It shows the probability how many times the sequence of w_1, w_2 appears in a corpus. For d-bigram, we add 'distance parameter' to bigram. D-bigram information takes the probability of a w_2 which appears in 'd'-distance away from a word w_1.

D-bigram includes mutual information and bigram. Figure 1 shows an image of relations among d-bigram, mutual information, bigram and trigram.

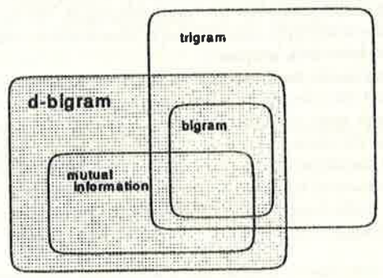


Figure 1: D-bigram and other informations

3.2 Mutual Information

D-bigram is similar to mutual information. Mutual Information in NLP [8] [9] is defined as follow;

$$MI(w_1; w_2) = log \frac{P(w_1, w_2)}{P(w_1)P(w_2)} \quad (1)$$

w_1, w_2 : a word
$P(w_i)$: the probability w_i comes out in a corpus
$P(w_1, w_2)$: the probability w_1 and w_2 comes out together in a corpus

This is an information of the association of two words. It does not have any concern about distance or sequence of two words. According to d-bigram, we can extend this mutual information as follow;

$$MI_d(w_1; w_2; d) = log \frac{P(w_1, w_2, d)}{P(w_1)P(w_2)} \quad (2)$$

P(w) : the possibility the word w appears in the corpus
$P(w_1,w_2,d)$: the possibility w_1 and $w_2$come out d words away from each other in the corpus

Now MI_d has a parameter d which is a distance of two words. Our system use MI_d for calculation.

4 Smoothing a Corpus

As described above, d-bigram information comes to be bias when the corpus is bias. For example, suppose we have a simple corpus like Figure 2.

```
He is a boy.
He is a student.
Ken is a boy.
```

Figure 2: Simple Corpus on the Bias

This example shows that word 'He' has an d-bigram information with the words 'boy' and 'student', however 'Ken' does not have any d-bigram information with the word 'student'. Considering about the meaning, it is possible that Ken is not a student indeed. However d-bigram can be used for syntactic analysis, so if we can have an information between 'Ken' and 'student' from this corpus, it is very effective. Our smoothing approach is based on this policy.

4.1 Relation among Words

To smooth a d-bigram information, we define a relation among words. When we take a d-bigram information from some corpus, we also get following informations;

N : the number of kinds of words in a corpus.
$c(w_i, w_j, d)$: frequency that word w_i appears at the distance of d words from word w_j.
$f(w)$: frequency that word w appears in the corpus.

Let W_i denote a feature matrix for some word w_i as expression (3).

$$W_i = \{c(w_i, w_j, d) | -N < i < N, -N < j < N\} \quad (3)$$

N : the number of kinds of word in a corpus.
w_i : The i-th word in the corpus
$c(w_i, w_j, d)$: frequency that word w_i appears at the distance of d words from word w_j.
d : distance between w_i and w_j.

This matrix will be characterized according to the words, and we use this to determine whether a certain word is resemble to another or not. We calculate the value of this likelihood with expression (4).

$$R(w_i; w_j) = \frac{W_i \cdot W_j}{|W_i| \cdot |W_j|} \quad (4)$$

This value denotes an approximate angle of feature matrices of two words. From this value we can measure how alike those two words are.

4.2 Getting a Cluster of Words

We can get clusters of words according to R_{ij}, the relation between the word w_i and the word w_j. Figure 3 shows an example of a cluster. Now we show the best 10 words of the cluster.

```
<< he >>          << masao's >>
  she 0.933650      junko's 0.800000
    i 0.860372     abraham 0.800000
   it 0.839873       kumi's 0.723196
   we 0.812462        ken's 0.655087
 they 0.793287            a 0.630961
 what 0.767187      mike's 0.605705
where 0.760661     raining 0.597614
 jane 0.754544      always 0.567367

<< baseball >>          << my >>
  tennis 0.734267          the 0.753352
basketball 0.569614    mike's 0.710319
 football 0.500712        his 0.663994
      not 0.447326        her 0.653838
        a 0.444852       this 0.643308
       to 0.443231       your 0.595691
       up 0.436360          a 0.588980
    still 0.425295       that 0.577314
  english 0.421694       very 0.576184

Points on the right shows a cos()
 according to an angle between two words.
```

Figure 3: Cluster of words

4.3 Smoothing a D-bigram

There were some research to smooth a statistical information for NLP. The major method to do so is based on a grammatical category of words[10]. To get the category of words, we need some grammatical information. Some corpus like Brown Corpus has tags, so we can get a category of the words easily. However it is difficult to make such corpus which has complete tag(category) information. In this paper we propose a

method which does not need any grammatical information but only statistical information. As we described in the previous section, we can get a cluster of words using d-bigram. This cluster shows that words, which are members of the cluster, have same characteristics in some ways, so we can give them a same value as a new feature matrix. Expression (6) is used to calculate the new value.

$$v_{tmp}(w) = \sum_{i=1}^{n} f_1(D_v(w, x_i)) f_2(frq(x_i)) v(x_i) \quad (5)$$

$$v_{new}(w) = \frac{|v(w)|}{|v_{tmp}(w)|} v_{tmp}(w) \quad (6)$$

x_i : The i-th word in the corpus
f_1 : the function which gives some weight to $D_v(w, x_i)$.
f_2 : the function which gives some weight to $frq(x_i)$.
$D_v(w_1, w_2)$: the function which gives distance between w_1 and w_2.
$frq(w)$: the function which gives frequency of word w.

5 Machine Translation

We have implemented statistics-based machine translation using d-bigram model. We have designed the system aiming at the multi-lingual translation, since statistics-based system is more suitable for it. We chose English, Japanese and Chinese for the system. We did not need any grammatical information for those languages, however we have to prepare the corpora and word-dictionaries. Figure 4 shows an image of multi-lingual system.

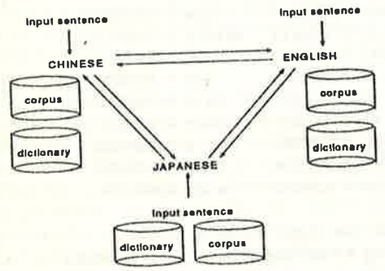


Figure 4: Image of Multi-Lingual System

5.1 Morphological Analysis

An English sentence is separated into words with blanks. So we can get a word very easily. However we can not get a word from both Japanese inputs and Chinese inputs, since both of them are agglutinative. For handling with agglutinative language, we have to analyze morphological information.

It is not so difficult to divide a sentence into some words by using word-dictionary, however we get a lot of answers as result. We have to have a method to select a correct answer at that point. We use mutual information, MI_d, to do it. We evaluate each answer using a point which is calculated with expression (7).

$$I_d(W) = \sum_{d=1}^{n-1} \sum_{i=0}^{n-d} \frac{MI_d(w_i, w_{i+d}, d)}{d^2} \quad (7)$$

W : a sentence
w_i : The i-th word in the sentence

We take the sentence which gets the highest point as the answer, then it is passed to the next step, the lexical translation.

5.2 Heuristic Approach to Lexical Translation

At this step, we translate a word of source language to a word of target language. We have word-dictionary for this. Word-dictionary has an entry for each word and a translation to a target language. Now we use heuristic approach to construct a dictionary.

Brown showed a method to make glossary in his paper[3]. His method is suitable for statistical model indeed, however we think it is hard to collect multi-lingual corpora for same domain. If we can do it, probably there is a word-dictionary for that domain. From this assumption we construct only word-dictionary for translation.

5.3 Creating a Result Sentence

A target sentence will be created at this step. We take a sentence which all words in the sentence are translated to words in a target language, however the order of those words are illegal for a target language. So we have to reconstruct the sentence according to a statistical information of a target language.

First, we get permutations of those words. For each elements of those permutation, we calculate the point of evaluation using expression (7). Then we take the sentence which gets the highest point as the answer.

We get too many permutations when a target sentence has many words. If we calculate all those sentences, it costs very much. So we cut some branches during calculation. It is important to get an answer in a reasonable period of time.

6 Results

We have tested our system under the condition describe below;

- Target languages are English, Japanese and Chinese.
- The corpus comes from English text book for junior high school students in Japan. For both Japanese and Chinese corpus, we translate all sentences manually.
- The corpus has 1407 kinds of words and 25000 words totally.

6.1 Word-Clustering

Table 1 shows an example that how relations between words can be gotten. We can get considerably good result in relations. 'he' and 'tom' get high association and 'boy' and 'student' too.

Table 1: Simple Relations of Words

	he	is	a	boy	student	tom
he	1.00	0.00	0.00	0.00	0.00	0.91
is	0.00	1.00	0.00	0.00	0.00	0.00
a	0.00	0.00	1.00	0.00	0.00	0.00
boy	0.00	0.00	0.00	1.00	0.94	0.00
student	0.00	0.00	0.00	0.94	1.00	0.00
tom	0.91	0.00	0.00	0.00	0.00	1.00

We can get clusters from these relations. Table 2 and Figure 3 shows examples of results. Though there are some noises in the result, elements of the cluster is considerably good. Especially, words 'tokyo', and 'japan, has words which decade places as their members. This cluster shows not only grammatical category but also semantical group of words. This is the very result we expected by using d-bigram.

Table 2: Clusters of Words

tokyo(19)		japan(49)	
japan	0.787259	america	0.805020
london	0.705708	tokyo	0.787259
america	0.698459	canada	0.764874
hawaii	0.688024	france	0.721230
canada	0.631273	london	0.719089
witzerland	0.627182	kyoto	0.716284
australia	0.623045	switzerland	0.714828
kyoto	0.598293	hawaii	0.714697
college	0.592640	college	0.669186
france	0.568182	australia	0.654228

6.2 Corpus-Smoothing

The result of smoothing is hard to show. Here we consider small corpus like Figure 5.

```
He is a boy.
He is a student.
He has a book.
He has a pen.
Tom is a boy.
```

Figure 5: Simple Example of a Corpus

From the corpus, we get a feature matrix of 'tom' as Table 3. After smoothing, we get new feature matrix as Table 4. From feature vector, 'he' is alike to 'tom', so smoothing increase the value, $c(tom, has, 1)$,$c(tom, boy, 3)$,$c(tom, student, 3)$,$c(tom, book, 3)$ and $c(tom, pen, 3)$. This result is quite good, and we can apply this method to translation or morphological analysis.

Table 3: the feature vector of "tom"(before smoothing)

tom	-3	-2	-1	1	2	3
he	0	0	0	0	0	0
is	0	0	0	1	0	0
a	0	0	0	0	1	0
boy	0	0	0	0	0	1
.	0	0	0	0	0	0
student	0	0	0	0	0	0
has	0	0	0	0	0	0
book	0	0	0	0	0	0
pen	0	0	0	0	0	0
tom	0	0	0	0	0	0

Table 4: the feature vector of "tom"(after smoothing)

tom	-3	-2	-1	1	2	3
he	0	0	0	0	0	0
is	0	0	0	0.6	0	0
a	0	0	0	0	1.21	0
boy	0	0	0	0	0	0.30
.	0	0	0	0	0	0
student	0	0	0	0	0	0.30
has	0	0	0	0.60	0	0
book	0	0	0	0	0	0.30
pen	0	0	0	0	0	0.30
tom	0	0	0	0	0	0

6.3 Morphology

Table 5 shows that most of the sentences, no matter whether the sentences are in the corpus or not, are segmented correctly. We find the right segmentation getting the best mark in the list of possible segmentations. We describe the details in another paper.

6.4 Translation

Table 6 shows the result of example of translations. This is the result of translating a Japanese sentence '彼はテレビを見る。(he watches the television)' into English and Chinese. The number on the right is a evaluated points calculated with d-bigram.

Now we can get 65-75% correct result with this style. We can say this is considerably good at this point. Most of errors occur in lexical analysis or morphological analysis. We can clear these errors to tune up our system, however some errors depends on semantical reason which we have to clear. Our system does not consider any semantical information now, so we have to add it in some way. For example, now we cna not distinguish the points of 'Tom loves Mary.' and 'Mary loves Tom.'. We have to consider a semantical information and even pragmatical or contextual information to solve this

Table 5: Result of Morphological Analysis

	the best point	the second best	the third best
α	100.0 %	100.0 %	100.0 %
β	100.0 %	100.0 %	100.0 %
γ	100.0 %	100.0 %	100.0 %
δ	95.0 %	100.0 %	100.0 %
ε	89.7 %	96.6 %	100.0 %

α : the very sentences in the corpus
β : replaced one morpheme in the sentence (the buried morpheme is in the corpus)
γ : replaced one morpheme in the sentence (the buried morpheme is not in the corpus)
δ : sentences not in the corpus (the morphemes are all in the corpus)
ε : sentences not in the corpus (include the morphemes not in the corpus)

problem. We think it is an interesting future work for our system.

Table 6: Example of Translation Result

彼はテレビを見る。

generated sentence	evaluation point
ENGLISH	
he watches the television	118.292
he watch the television	104.761
he sees the television	103.458
CHINESE	
他 看 電視	70.599
他 看見 電視	46.455
他 見 電視	22.016

7 Conclusion

Our statistical model, d-bigram, turned out to be very efficient in NLP. Clusters of words and smoothing method using the clusters is quite meaningful for self-organizing of corpus. And also, d-bigram is useful for the morphological analysis. Its result for agglutinative-language is absolutely good, there is no doubt that d-bigram is a powerful method in segmenting agglutinative-language sentences.

Using d-bigram to multi-lingual machine translation shows pretty good result(60%). This percentage is not so bad as a result of multi-lingual machine translation, so we could show the possibility of d-bigram model to use for multi-lingual machine translation. We have to clear a few problems to increase this percentage hereafter. The most important problem is lexical analysis which has any semantical or pragmatical information. Most of errors occur at the point in our system.

References

[1] M Nagao. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. *Artificial and Human Intelligence*, pp. 173–180, 1984.

[2] S. Sato and M. Nagao. Toward Memory-based Translation. *Proc. of COLING-90, Vol. 3*, pp. 247–252, 1990.

[3] et al. Brown, P. A Statistical Approach to Machine Translation. *Computational linguistics Vol.16, No.2*, pp. 79–85, 1990.

[4] et al. Kaji, H. Learning Translation Templates from Bilinguial Text. *Proc. of the COLING '92*, pp. 672–678, 1992.

[5] J.D. Ferguson. Hidden Markov Models for Speech. *Princeton, New Jersey, IDA-CRD, Oct.*, pp. 8–15, 1980.

[6] J. Tsutsumi, T. Nitta, K. Ono, and S. Nobesawa. A Multi-Lingual Translation System Based on A Statistical Model(written in Japanese). *JSAI Technical report, SIG-PPAI-9302-2*, pp. 7–12, 1993.

[7] F. Jelinek, R. Mercer, and S. Roukos. Classifying Words For Improved Statistical Language Models. *IEEE*, pp. 621–625, 1990.

[8] K Church, W. Gale, P. Hanks, and D Hindle. Parsing, Word Associations and Typical Predicate-Argument Relations. *International Parsing Workshop*, 1989.

[9] D. Magerman and M Marcus. Parsing a Natural Language Using Mutual Information Statistics. *AAAI*, 1990.

[10] K. Shikano. Improvement of Word Recognition Results by Trigram Models. *IEEE ICASSP*, pp. 1261–1264, 1987.

INFORMATIONAL MEASURES OF CAUSALITY

Juhan Tuldava
Tartu, Estonia

1. Causality, in general, is understood to involve the effect of one event, process, or entity upon another. It is supposed to be the necessary connection of events through *cause* and *effect*. The main issue in the discussion of causality in our day is the question of whether scientific investigation should proceed on the assumption that things are definitely determined by their "causes" or on the understanding that it is merely probable that one thing "flows from another" (cf. Dreher 1983). The last statement leads quite naturally to the probabilistic conception of causality, according to which "causality is something that may be found to a greater or smaller degree and not only exist" (Wiener 1956). Using this approach it is assumed that some event (process, entity) may be the cause of some other if the appearance of the first (e.g. *X*) with a high degree of probability is followed by the appearance of the other (e.g. *Y*) while it is stated that, in symbolic form, $P(X) > 0$ and $P(Y/X) > P(Y)$, i.e. "the appearance of *Y*, on condition that *X* appeared, is more probable than the appearance of *Y* without *X*". In case of multiple causes it turns out that $P(Y/X_1, X_2 ...) > P(Y/X_1)$, i.e. the addition of new factors increases the probability of the appearance of *Y*.

According to the probabilistic approach, deterministic causality is naturally included in the probabilistic scheme of causality as a particular case which has the probability equal to *1*. Because of the fact that chance and secondary factors cannot be excluded from the interrelations between events, dependence necessarily acquires a probabilistic (stochastic) character.

In the functioning of language a complicated set of multiple mutually interrelated features is operating, However, in some cases the dominating influence of one or two factors can be established. This problem of clarifying causality is to be handled in each concrete case in a way congruent with the aims and tasks of the investigation, as well as its professional context.

2. Causal analysis in its probabilistic treatment can be considered one of the most important subsidiary methods for the description and explanation of complex systems. The problem lies in the investigation of the possibilities of an adequate measurement of causality. It calls for the use of various quantitative methods.

(i) When using traditional *statistical methods*, the dependence of an event *Y* on another event *X* may be established and measured with the help of the

methods of simple, partial and multiple correlation and regression (see, e.g., Tuldava 1994).

(ii) A more complicated method available which can help scientists establish the relative importance of different sources of causality by "partitioning" causality between several predictive variables (factors) is the so-called *path analysis* (see, e.g. Heise 1975). This method (and a number of its modifications) is based on linear dependence and it may be considered a special case of regression analysis where the regression coefficients are interpreted in terms of causal relations.

(iii) Causal relations between events are manifest not only in correlations between the various states of events, but also in the dependence between the levels of "uncertainty" (entropy) in a given system. This leads us to *informational measures* of causality which will be the main topic in this report.

3. We shall examine the simplest case, when a binary causal relation between two variables X and Y is involved. In terms of information theory we can speak of a causal dependence between X and Y if knowledge of X is capable of reducing the uncertainty of Y, or vice versa. In other words, to measure a dependence between X and Y means to measure the amount of reduced uncertainty of Y (or X) within the system (X,Y) against the amount of uncertainty of Y (or X) without considering the system (X,Y). As statistical measures of uncertainty, the formulas for entropy and information are used (cf. Kullback 1959):

$$H(X) = - \sum_i p_{i.} \ln p_{i.} ; \qquad (1)$$

$$H(Y) = - \sum_j p_{.j} \ln p_{.j} ; \qquad (2)$$

$$H(X,Y) = -\sum_i \sum_j p_{ij} \ln p_{ij} ; \qquad (3)$$

$$I(X,Y) = -\sum_i \sum_j p_{ij} \ln (p_{ij}/p_{i.}\, p_{.j}) = H(X) + H(Y) - H(X,Y). \qquad (4)$$

Here p_{ij}, $p_{i.}$ and $p_{.j}$ are the probabilities (see Table 1), $H(X)$ and $H(Y)$ are the entropies of X and Y when considered independently, $H(X,Y)$ - entropy of the joint event, and $I(X,Y)$ - the amount of "shared information" which serves as a *symmetrical measure* of the intensity of dependence between X and Y. The relations between these measures are graphically presented in Fig. 1.

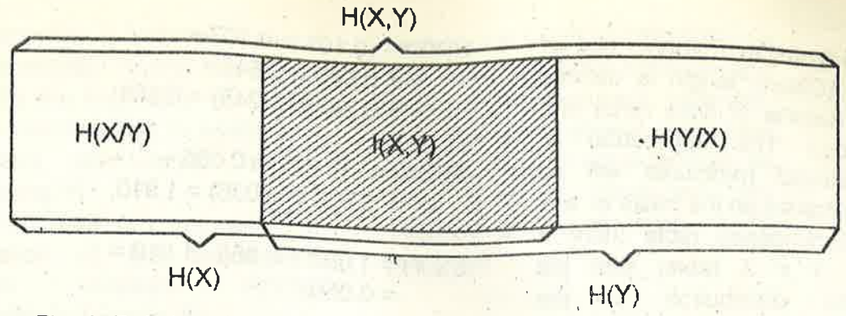


Fig. 1. Relations between information and entropy within the system of two variables X and Y.

The informational measures based on *conditional entropy* of X or Y are of special interest (cf. Fig. 1):

$$H(X/Y) = H(X,Y) - H(Y) \qquad (5)$$

measures the entropy of X when Y is known, and

$$H(Y/X) = H(X,Y) - H(X) \qquad (6)$$

measures the entropy of Y when X is known.

On the basis of conditional entropies, the following *standardized asymmetric* measures of dependence can be constructed:

$$R(X/Y) = [H(X) - H(X/Y)] /H(X) = I(X,Y)/H(X) \qquad (7)$$

and

$$R(Y/X) = [H(Y) - H(Y/X)] /H(Y) = I(X,Y)/H(Y). \qquad (8)$$

The measurement of causal dependence between X and Y according to formulas (7) and (8) is the estimation of the relative amount of reduced uncertainty of one the variables within the system of (X,Y) against the amount of uncertainty of the variable considered independently. In other words, the coefficients signify the diminishing level of uncertainty of one of the variables under the influence of the other variable. The coefficients vary within the limits of *0* and *1*.

4. An example where causality in linguistics has been ascertained is in the case of *Menzerath-Altmann's law* stating: "The longer a language construct the shorter its components (constituents)" (Altmann & Schwibbe 1989). Or, otherwise formulated: "The length of the components is a function of the length of language constructs".

As an illustration we shall analyze the dependence of average word length (in syllables) on clause length (in words) in a

sample from an Estonian text of fiction. (Clause length is defined by the number of finite verbs in a sentence.) The application of informational measures will be demonstrated on the basis of a *R x C* contingency table (here a simple 3 x 3 table) with the bivariate distribution of the features "clause length" *(X)* and "word length" *(Y)*; see Table 1.

Table 1
Absolute and relative frequencies of cross-classified data

	Y_1	Y_2	Y_3	Total
X_1	2 (p_{11} = 0.005)	87 (p_{12} = 0.207	54 (p_{13} = 0.129)	143 ($p_{1.}$ = 0.341)
X_2	23 (p_{21} = 0.055)	100 (p_{22} = 0.238)	35 (p_{23} = 0.083	158 ($p_{2.}$ = 0.376)
X_3	20 (p_{31} = 0.047)	84 (p_{32} = 0.200)	15 (p_{33} = 0.036)	119 ($p_{3.}$ = 0.283)
Total	45 ($p_{.1}$ = 0.107)	271 ($p_{.2}$ = 0.645)	104 ($p_{.3}$ = 0.248)	n = 420 ($p_{..}$ = 1.0)

In table 1, X_1 denotes short clauses (less than 5 words), X_2 - medium long (5-8 words), and X_3 - long clauses (more than 8 words); Y_1 denotes short words (average length less than 2 syllables), Y_2 - medium long (2-2.5 syllables), and Y_3 - long words (more than 2.5 syllables).

The calculation of the informational measures with the help of formulas (1-4) gives the result:

$$H(X) = - (0.341 \ln 0.341 + 0.376 \ln 0.376 + 0.283 \ln 0.283) = 1.092;$$

$$H(Y) = - (0.107 \ln 0.107 + 0.645 \ln 0.645 + 0.248 \ln 0.248) = 0.868;$$

$$H(X,Y) = - (0.005 \ln 0.005 + \ldots + 0.036 \ln 0.036) = 1.910;$$

$$I(X,Y) = 1.092 + 0.868 - 1.910 = 0.050.$$

To test the statistical significance of the computed value of $I(X,Y)$, i.e. the significance of interdependence (association) between *X* and *Y*, the loglikelihood ratio statistic G^2 can be used:

$$G^2 = 2 n I = 2 n [- \Sigma_i \Sigma_j p_{ij} \ln (p_{ij} / p_{i.} p_{.j})]. \quad (9)$$

In our case $G^2 = 2(420)0.050 = 42.00$. As is known, the statistical tests G^2 and X^2 are of similar functioning. Both have a distribution which is approximately chi-squared with *(r-1)(c-1)* degrees of freedom. Applied to the 3 x 3 table, df = 4 and the critical value at the 0.001 level is 18.46. Consequently, the value of G^2 = 42.0 allows us to state that the

interdependence between *X* and *Y* is statistically highly significant (despite the low value of $I(X,Y)$, which indicates that in addition to *X* there are other factors influencing *Y*).

We can proceed with the calculation of the asymmetric relative measure $R(Y/X)$ according to formula (8):

$$R(Y/X) = 0.050/0.868 = 0.058.$$

The value of this index signifies the diminishing level of uncertainty of *Y* under the influence of *X*. As Astola & Virtanen (1983) have pointed out, a square root transformation of this index would be a better measure of dependence which would fulfil both theoretical requirements and intuitive expectation set for a correlation coefficient. We calculate

$$\sqrt{R(Y/X)} = \varphi(Y/X) = \sqrt{0.058} = 0.241$$

which gives us the estimated degree of causal determination of *Y* (average word length) by *X* (clause length).

Note. The concrete values of informational indices may vary depending on the grouping of categories in an *R x C* table. In this study a somewhat simplified grouping of clause length and word length into 3 subcategories was used.

As informational measures of causality do not require linearity in the relation between the variables and can be calculated on the basis of both metrical and non-metrical (nominal) variables, they (i.e. informational measures) deserve special consideration as valid tools in measuring and interpreting causal relations in language.

References

Altmann, G., Schwibbe, M.H. et al. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim etc., Olms.

Astola, J. & Virtanen, I. (1983). *A measure of overall statistical dependence based on the entropy concept*. Vaasa, Univ. Press.

Dreher, J.P. (1983). Causality. In: *Encycl. Amer.*, vol. 6.

Heise, D.R. (1975). *Causal analysis*. New York, Wiley.

Kullback, S. (1959). *Information theory and statistics*. New York, Wiley.

Tuldava, J. (1994). *Methods in quantitative linguistics*. Trier, WVT (to appear).

Wiener, N. (1956). *I am a Mathematician*. Garden City, N.Y., Doubleday.

Typological and Stylistic Characteristics of the Phonetic Word with Examples from some Indo-European Language

Zlatoustova L.V.,
Moscow State University, Philological Faculty,
Laboratory of Phonetics and Speech Communication
Phone: (095) 939-32-56

Topical paper

AREA: descriptive quantitative typology and stylistics of languages in phonetic aspect

SUMMARY:

In the present paper the distribution of the most frequent phonetic words in a number of Indo-European languages is considered. Also a method of obtaining represetative statistic data is described.

The system of prosodic features is common to all phonostylistic varieties of oral speech. This includes such subsystems as rhythmic structures of phonetic words, syntagmas and their rhythmic models, phrases, phonoparagraphs. However the distribution of units in each of the subsystems, their frequencies of occurence and probabilities are specific for different functional styles.

It is evident that units of grammatical and lexical levels of this or that functional style crucially influence the choice of prosodic units. A certain phonetic style is formed where linguistic means, communicative set and a number of extraliguistic conditions meet. The set for speech realization is of great importance, each phonostyle has it's primary functional significance pertaining to the specific character of style.

The typology of phonetic words and rhythmic structures (RS) is a system fundamental for forming the prosody of units of different length where RSs are construed as one or more words united by one word-stress. RS-type is a fraction where the number of syllables is the numerator and the ordinal number of the stressed syllable is the denominator, e.g. год, table: RS - 1/1; год/а, the t/able: RS - 2/2; г/оды, r/ecord: RS - 2/1.

Quite a number of articles are devoted to research of word-stress in Indo-European languages. However there is no agreement on characteristics and functions of the phenomenon in question. In the present paper we assume that the primary function of the word-stress is to organize the phonetic word but not to highlight the stressed syllable.

In each of the studied languages the grammatical properties determining the rules for forming RSs are taken into consideration, namely: presense vs absence of proclitics and enclitics, combining of autonomous word into RSs, gemination, junctures, rules of reduction of various types, etc.

The following inventory of frequent RSs (tables 1, 2 and the figure) is a product of segmentation of recorded texts into RSs by a qualified group of listeners. The segmentation was carried out against the background noise; the semantic content was not recognized but the rhythmic structure could be restored. The white noise was generated in the range 20-20.000 Hz.

RSs form the rhythmic framework of a syntagma. We call this feature "rhythmic framework of a syntagma" (RFS). It is this framework that "holds" the melody of a syntagma based on the first and the last syllables of RFS.

Including the melodic parameter in the analysis results in a new unit called rhythmic and melodic framework of a syntagma.

Table I.

Rhythmic structures

RS type %	Languages				
	Russian	Bulgarian	Spanish	English	German
1/1/	13	13,8	5,8	27	17,6
2/1	16,8	16,5	18,3	18,8	16,9
2/2	21,3	10,5	7	21,5	16,7
3/1	6	7,8	1,3	1,4	7
3/2	19,6	16,6	23,3	5,5	23,3
3/3	7	2,3	6,7	3,8	3,2
4/1	1,8	1,8	-	0,7	-
4,2	4	10,9	1,7	1,9	2
4,3	9,2	11,7	20	1,9	2,8
4/4	1,4	1,3	0,3	0,5	0,5
5/3	5	2	5	1,4	1
5/4	-	-	8,3	-	-

Table 2.

Rhythmic structures in various functional styles of Russian speech

Structure types %	Prosaic texts			
	Fiction and journalism	Scientific text		Political text
		linguistic !	law	
1	5,28	11,74	9,49	7,73
2/1	12,4	15,93	7,29	12,64
2/2	14,86	11,19	7,56	8,59
3/1	4,74	6,91	11,04	6,74
3/2	18,11	13,02	6,43	12,52
3/3	7,83	5,96	8,45	11,29
4/2	8,71	8,82	10,59	8,2
4/3	8,18	6,06	10,92	6,2
5,3	7,48	7,58	5,98	4,96
% of text covering	87,59	81,21	76,73	78,60

The material of the tables and figure reveal a common tendency of using the most frequent structures in studied

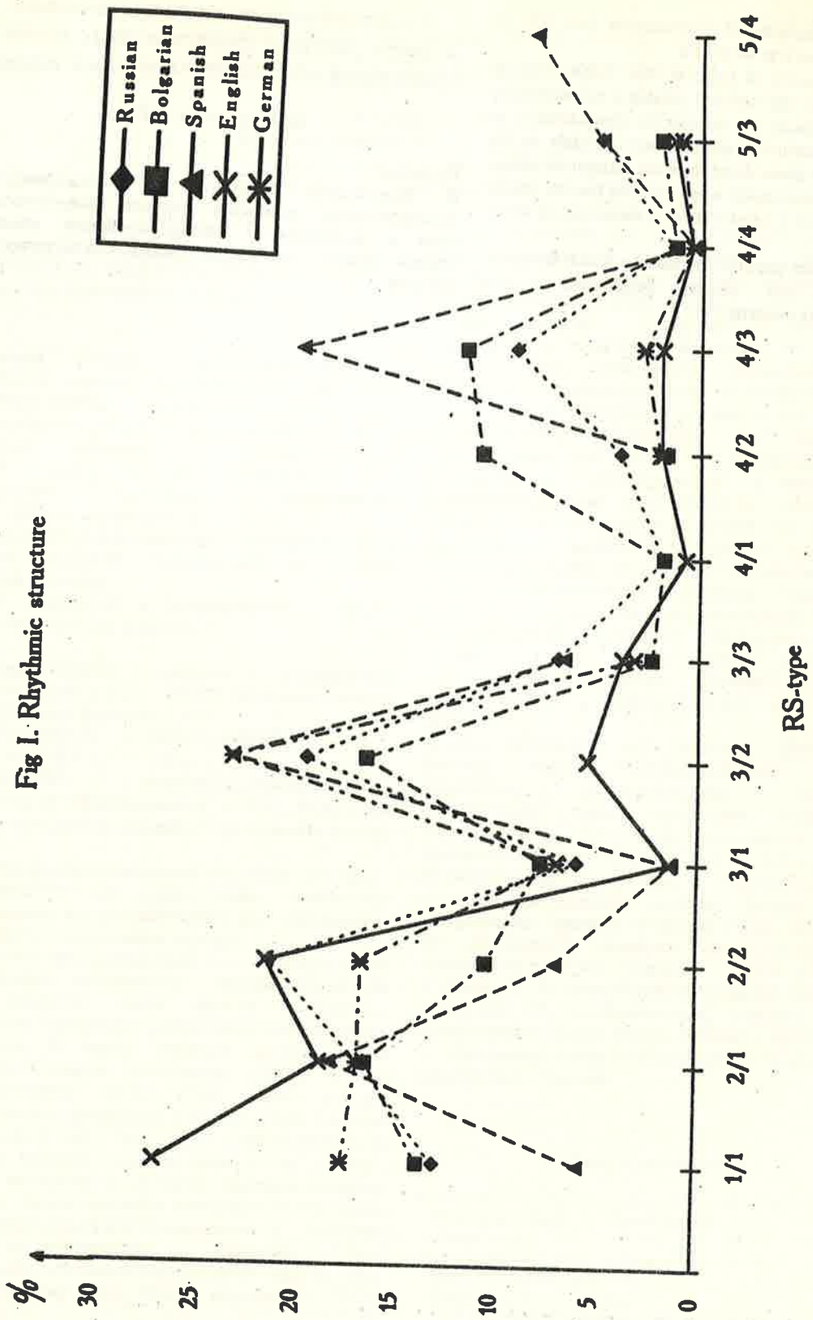


Fig I. Rhythmic structure

languages. The reliability of segmentation into RSs by native speakers was no less than 93%.

The limited inventory of frequent RSs helps indetify each type even when the stressed syllable is not marked by any attribute since there are no specific characteristics of stress. Only RSs cosntituting special neuron models in the human memory in a generalized form are recognized. RS as a unit of the expression aspect when realized has the lexico-semantic meaning of a word-form or a succession of word-forms.

This approach makes possible to describe lexico-semantic, semantico-syntactic and prosodic peculiarities of text organization on their integrity.

Типологические и стилистические характеристики фонетического слова с примерами из индоевропейских языков

Златоустова Л.В.

Резюме:
В настоящем докладе рассматривается распределение наиболее частых фонетических слов в некоторых индоевропейских языках. Также описан метод получения статистических данных.

Вероятностно-алгоритмическая модель порождения русского стихотворного текста

Зубов А.В.,
Минский лингвистический университет
220662 г.Минск, ул. Захарова, 21
E-mail: rootalingua@.ibibel..glas..apc..org

Доклад

ТЕМАТИЧЕСКАЯ ОБЛАСТЬ: Моделирование на ЭВМ процесса порождения текста.

Резюме:
Предлагается модель порождения русского стихотворного текста, реализованная на IBM-совместимых ПЭВМ. Модель опирается на базу знаний, представленную в виде фреймов, и на специальный словарь, в котором каждой машинной основе наряду с грамматической и семантической приписана информация эмоциональная и структурная.

Актуальность темы связана с необходимостью разработки моделей порождения на ПЭВМ письменных текстов.

Работа относится к направлению - синтез текстов естественных языков.

В предлагаемой модели порождения будущее стихотворение VERS представляется в виде следующего фрейма:

VERS[<SESFOS >;<RO1>;<SESFOS >; ... <SESFOS >;<RK>], (1)

где SESFOSi - семантико-синтаксические формулы строф стихотворения, а RO1, RO2, ... Rk - правила, которые связывают эти строфы между собой.

Семантико-синтаксическая формула строфы - это записанное на специальном семантико-синтаксическом языке СЕМСИНТ [1] предметно-эмоциональное содержание строфы.

Предметно-эмоциональным содержанием (ПЭС) строфы называется опирающееся на главные опорные слова текста сообщение, раскрываемое детально одной или несколькими микротемами. К числу главных спорных слов стихотворного текста относятся: ГОС1 - 1-ое главное опорное слово ("я", "мы" и их контекстуальные синонимы); ГОС2 - 2-ое главное опорное слово ("ты", "вы", "Вы" и их синонимы); ГОС3 - 3-е главное опорное слово ("он", "она", "они" и их синонимы). К числу главных опорных слов текста будем относить и те существительные (их контекстуальными синонимами), которые имеют первостепенное значение для формирования основного содержания текста. Эти главные опорные слова будем обозначать ГОС4, ГОС5 и т.д.

Общее предметно-эмоциональное содержание строф автора выявляется путем качественного анализа стихотворений этого автора, и оно записывается словами естественного языка. Например, общее предметно-эмоциональное содержание строфы (в модели оно имеет код S29) может быть представлено так: "Констатация некоторого обращения к С2, некоторого состояния и/или действия С1, связанного с С4". В анализируемом множестве стихотворений выделено 74 типа таких строф (они обозначаются SO1, SO2, ... S74).

Еще одной важной особенностью поэтического текста является то, что его семантическая связность определяется как результат взаимодействия компонентов такого текста как по горизонтали, так и по вертикали. Семантическая связность по горизонтали, специфичная для любого прозаического текста, регулируется синтаксической и семантической валентностью слов предложения. семантическая связность поэтического текста по вертикали предполагает наличие у группы слов (или нескольких групп слов) такого текста какого-то общего семантического признака. Такая группа слов образует единую семантическую цепочку - семантическую изотопию. Например, слова "вечер", "утро", "час", "секунда" имеют общий синтаксический признак "время" и поэтому образуют одну семантическую изотопическую цепочку.

Составляющие таких цепочек, повторяясь в разных местах стихотворения, являются средством развертывания содержания поэтического текста, средством создания его определенного образно-эмоционального фона. В конкретном стихотворении может быть несколько изотопических цепочек. Множество слов, входящих в каждую цепочку, будем называть микротемой. Главные опорные слова текста и все его микротемы образуют тему стихотворного текста (от др.-греч. "thema" - "нечто, положенное в основу"). В предлагаемой модели порождения выделено 77 изотопических цепочек, которые обозначаются через МОО1, МОО2, ... МО77.

Например, тема стихотворного текста задана следующим образом:

Табл. 1

Тема стихотворения

Тип слов темы	Код	Слова темы
ГОС1	С1/1	мы
ГОС2	С2/1	родная
	С2/2	милая
ГОС4	С4/1	день
	С4/2	миг
	С4/3	полночь
Микротемы	МО24	карнавал, веселье, песни, маски
	МО26	снег, метель
	МО27	жизнь, сказка

В такой таблице после наклонной линии указываются номера синонимичных (в широком плане) опорных слов.

В состав семантико-синтаксического языка СЕМИНТ входят: 20 семантических функций, подобных семантическим падежам Ч.Филлмора [2], однозначные коды предлогов, местоимений, союзов и союзных слов, частиц. Сюда же входят коды семантических подклассов существительных, прилагательных, глаголов, причастий, деепричастий, наречий и междометий [3, 143-162].

Например, семантико-синтаксическая формула строфы, реализующая упомянутое выше предметно-эмоиональное содержание S29, связанная с только что указанной темой стихотворения, и состоящая из 4-х строк запишется так:

AAG<NO36**C2/1,NO36**C2/2>!AEL<P19*NO67**C4/1>+R1/2<VO34>!

R4/1<T21*VO42>+AS2<PO5>+AP2<PO2**C1/1>+AO4 <T25*NO96**26>;

AO4

<T25*NO96**26>

CO5+R1/2<VO34>+AS2<P11*AOO4*NO45**24>,

CO1+R3/4<VOO2>+AO1<NO61>+AB2<T25*NO88**24>,AB2<T25*NO45**24>!

В этой формуле: AAG, AEL, AS2, AP4, AB2, AO4 - семантические функции языка СЕМСИНТ, NO36, NO67 ..., VO34, VO42, ..., AOO4, A01 - коды семантических подклассов соответственно существительных, глаголов и прилагательных; P19, PO5, PO2, T24, T25, CO5, CO1 - однозначные коды соответственно местоимений ("тот", "он", "нас"), частиц ("не", "ни"), союзов ("но", "и"). Через C1/1, C2/1, C2/2 и C4/1 показано место главных опорных слов. Вместо кодов изотопических цепочек MO26, MO24 здесь указываются только две последние цифры. Наконец, через R1/2, R4/1, R3/1 в формуле указано место сказуемого в предложении. Первая цифра после R обозначает управление глагола, а цифра после "/" - время глагола (1 - настоящее, 2 - прошедшее).

Если говорить о правилах взаимного расположения таких строф, то необходимо отметить следующее. В прозаическом тексте порядок расположения обзацев зависит от общего замысла произведения, его фабулы, необходимости развития действия в желаемую для автора сторону. В относительно небольших поэтических произведениях дело обстоит несколько иначе. Как отмечает В.Маяковский [4] и другие поэты и исследователи поэтического творчества, архитектоника стиха (взаимное расположение строф) определяется в основном интуицией автора, желанием достичь определенных эмоциональных воздействий. Такой механизм выбора и расстановки строф может быть смоделирован путем подключения в систему порождения стихотворного текста датчика случайных чисел.

Вместе с тем анализ большого числа стихотворений показал, что не все строфы могут быть выбраны случайным образом. Некоторые из них непосредственно связаны с предыдущими. Такая взаимосвязь строф во фрейме VERS обозначена фреймами низшего уровня с именами RO1, RO2 ... Rk. В общем случае фрейм Rk выглядит следующим образом:

Rk [<RSk>; <RFk>],

где RSk - чисто содержательные правила, ограничивающие порядок следования строф с определенными общими предметно-эмоциональными содержниями. Чаще всего сюда относятся слчаи, когда в последующей строфе употреблены контекстуальные синонимы главных опорных слов или же дается характеристика главного действующего лица или его действия без упоминания его имени. Во всех подобных случаях нет формальных показателей связи между строфами. Такие виды правил называются содержательными и формализовать их пока невозможно. Гораздо чаще распространены правила второго рода - RFk. Это - формальные правила, ограничивающие порядок взаимного расположения строф, записанных на языке СЕМСИНТ. Сюда относятся всевозможные логико-смысловые скрены, которыми начинаются первые строки последующих строф, наличие в первой строке последующей строфы кавычек, многоточий и т.д.

В процессе создания стихотворения в соответствии с формулой (1) семантико-семантические формулы строф ищутся в базе знаний в файле STROPH:

STROPH[<CODLXi>;<TEi>].

В нем CODLXi - имя фрейма младшего уровня, содержащего перечень кодов главных опорных слов, с которыи может употребляться заданное во фрейме TEi предметно-эмоциональное содержание строфы. Например, фрейм CODLXi для строфы с приведенным выше предметно-эоциональным содержанием S29 запишется так: CODLXi [<C1,C2,C4>].

В свою очередь фрейм TEi состоит из фреймов более низкого уровня:

TE[<Si>;<SESFOS ,ISC >;<SESFOS ,ISC >...],

где Si - общее предметно-эмоциональное содержание строфы; SESFOSi -семантико-синтаксические формулы строфы, которыми можно реализовать общее предметно-эмоциональное содержание Si и которые записаны на языке СЕМСИНТ; ISCi - имя фрейма низшего уровня, содержащего коды изотопических цепочек, специфичных для каждой семантико-синтаксической формулы.

Так, для уже упоминаемой строфы с предметно-эмоциональным содержанием S29 фрейм STROPH будет выглядить так:

STROPH[<C1,C2,C4,>;<S29>;<SESFOS ,MO24,MO26>;
SESFOS ,MO34>;<SESFOS ,MOO2,MO59>;...],

Как видно, каждое общее предметно-эмоциональное содержание может быть

представлено несколькими семантико-сентаксическими формулами на языке СЕМСИНТ.

Для порождения стихотворения задается тема текста, число К строф в стихотворении, ритм стиха и тип рифмы. Затем по датчику случайных чисел в базе знаний STROPH находится строфа с некоторыми ПЭС и выясняется, может ли она быть первой строфой стихотворения. Если может, уточняется, все ли ГОС, входящие во фрейм Si есть в теме стихотворения. Далее среди формул SESFOSi, соответствующих строке с данным ПЭС, случайным образом выбирается одна из формул SESFOSk. После этого выясняется, есть ли в выбранной формуле соответствующие теме текста изото пические цепочки, содержащиеся во фрейде ISCk. Если есть, семантико-синтаксическая формула первой строфы считается найденной. Затем аналогичным образом ищутся другие К-1 формул. При этом каждый раз по упомянутым выше правилам файла RFk проверяется, может ли стоять формула строфы N+1 после формулы строфы N.

Когда полностью построена семантико-синтаксическая формула всего стихотворения, производится заполнение ее словами из специального словаря, являющегося частью базы знаний системы порождения стихотворного текста. В таком словаре каждой квазиоснове кроме грамматической и семантической приписаны эмоциональная и структурная информация. последняя и используется для построения необходимого ритма строки и подбора рифмующихся слов.

На последнем шаге порождения производится морфологическое оформление слов предложений, входящих в строфы. Основой для этого служит структура предложений и правила языка СЕМСИНТ.

Литература

1. Зубов А.В. Вероятностно-алгоритмическая модель порождения текста (семантико-синтаксический аспект). Дис. ... докт. филол. наук. - М., 1985.

2. Филлмор Ч. Дело о падеже // Новое в зарубежной лингвистике. Вып.X. Лингвистическая семантика. - М.: Прогресс, 1981.

3. Зубов А.В., Зубова И.И. Основы лингвистической информатики. Часть 3. Исскуственный интелект. - Минск: МГПИИЯ, 1993. - с. 204.

4. Маяковский В. Как делать стихи // Маяковский В.В. Сочинения в двух томах. Том второй. Поэмы. Пьесы. Проза. - М.: Правда, 1988, с. 664-697.

Probabilistically - Algorithmic Model of Russian Verse Text Generation

Zubov A.V.

Summary:
Some model of Russian verse text generation is present. It is implemented on IBM-type PCs. The model is based on a knowlege base, realized by way of frame ideology and a specialized dictionary. The dictionary contains word-steams characterized not only grammatically, but also emotionally and structurally.

Establishment of Differential-Characteristics of Inductive Classes by Means of Discriminant Analysis (on the Material of Poetic Works by English Romanticists)

Andreyev S.,
Smolenskz Pedagogical State Institute,
Foreign Languages Faculty
Russia, 214000 Smolensk, Przhevalsky st, 4
Phone: (081-22)-37700

Vislinskaya E.
Smolenskz Pedagogical State Institute,
Foreign Languages Faculty
Russia, 214000 Smolensk, Przhevalsky st, 4
Phone: (081-22)-37700

Topical paper

AREA: Discriminant analysis of literary texts

Summary:
An attempt has been made to use discriminant analysis in a comparative study of the formal characteristics of works by English poets-romanticists: Byron, Wordsworth and Keats. Thirty lyric verses of each author were analysed.

There are two principal approaches to the classification of units which can conventionally be called "deductive" and"inductive".

D e d u c t i v e approach consists in finding an optimum devision of the investigated objects into groups depending on the level of similiarity or "likeness". The procedure is usually based on some formal method of finding similiarity between objects or groups of objects. The resulting number of classes or nature of the classes, received as a result of such analysis, in most cases can not be predicted in advance and is subject to "interpretation of the results"- an important final stage of "deductive" classification.

In i n d u c t i v e classifications the classified objects are grouped on the basis of previous research and intuition of the investigator. This approach was widely used in biology where it goes back to the taxonomy of K. Linne' according to whose aphorism the genus determines the characterisitcs and not vice versa. In other words in inductive approach the classes themselves at this stage of the analysis already exist, the purpose of classification consists in establishing differentiating and integrating characteristics for the empirical classes and in finding to what of the existing classes "difficult cases" (objects which fall into more than one class) refer. These two main aims of inductive classifications are often referred to as the problem of "discrimination of taxons", or "interpretation of classification", and the problem of identification of objects.

Thus, interpratation of classification in the deductive approach consists in establishing the most appropriate number of classes and comparing them with already existing notions in the science, whereas in the inductive approach it means establishing such characteristics of the classified objects which will then help to discriminate between the classes.

Allowing the necessary objectivity of the deductive classifications, we still find it necessary to point out that most of the classes in linguistics, as well as in stylistics, literary criticism and other branches of philology are, certainly, inductive.

One of the most appropiate methods of investigating the inductive classes for the purpose of solving the above-mentioned problems is discriminant analysis - a statistical method which allows one to study the differences between two or more groups of objects by several variables simultaneously and provides a method for predicting a small number of discriminating variables.

The application of this method has received interesting results in quite a number of investigations in biology, anthropology, psychology, and other social sciences, etc. At the same time it should be noted that in linguistics discriminant analysis has not been widely used.

In this investigation an attempt has been made to use discriminant analysis in a comparative study of the formal characteristics of works by English poets-romanticists: Byron, Wordsworth and Keats. Thirty lyric verses of each author were analysed. The poems were selected randomly from collections of poems of these authors. Lyric verses were chosen whose length does not exceed fifty lines. The total amount of lines for Byron equals 686, for Wordsworth 637, for Keats 516.

The scheme of characterisics suggested by V.S.Bayevsky (1 for measuring the lyrics was used. It includes such characteristics as "the type of strophe", "changeability of rhythm", "syntactic expressiveness", "inversion", "relative length of sentences", "intonational unity of the verse", etc.

The material for discriminant analysis thus consists of three classes of lyrics (30 each). Nine characteristics (variables) were used for their measurements, done on every object (lyric). The aim of the analysis consists in finding differential features, and more generally in estimating the degree of difference between the groups. This information will help answer the question: Which tendencies, integral or differentiating are reflected in the works of the authors? The presence of the first tendency can be explained by the fact that all three poets belong to one and the same literary trend, the second tendency, if discovered, should be ascribed to their significant divergence within it.

Technically discriminant analysis was carried out with the aid of the "STATGRAPH" computer program. The following main results were achieved.

Two canonical discriminant functions were generated. Using the standardized discriminant coefficients, it is possible to determine the relative importance of different characteristics.

The greatest contribution to the first function was made by the variables "presence of polimetric composition", "expressive syntax", "small relative length of the sentence". In the second function dominates the variable "type of strophical organization", then in relative importance come two variables: "absence of syntactic expressiveness" and "large relative length of the sentence". The first function comprises 68 per cent of the total discriminant strength, the second 32 per cent. The coefficient of canonical correlation which reflects the degree of dependence between classes and descriminant functions has the following meanings: 0.52 for the first funcion, 0.32 for the second. Thus, the greater amount of information is given by the first function. Both functions are statistically significant.

Table 1 shows the location of group centroids for the unstandardized discriminant functions. It must be noted that the first function (column 1) discriminates rather well between the objects of class 1 (Byron's lyrics) and the objects of other classes. Function 2 differentiates class 2 and class 3.

TABLE 1

Group centroids		
1	-0.83846	-0.04164
2	0.36713	0.52346
3	0.47133	-0.48181

The actual and predicted classification results are the following: only 55 per cent of all the objects coincide with the predicted classification, but 70 per cent of them are Byron's works. Thus, from the point of view of formal characteristics Byron's lyrics possess a stronger differential force within the romantic literary trend, while the lyrics of Keats and Wordsworth,. in spite of some temporal and ideological differencies, are marked by integral tendencies.

The report also contains the comparison of discriminant and cluster analysis and the comparison of romantic lyrics with the works of other literary schools.

Установление дифференциальных характеристик индуктивных классов на основе дискриминантного анализа

Андреев С., Вислинская Е.

Резюме:
Осуществлена попытка применить дискриминантный анализ в сравнительном исследовании формальных характеристик произведений английских поэтов-романтиков: Байрона, Вордсворта и Китса. Проанализировано по 30 лирических стихотворений каждого автора.

Length of a Chinese Word in Relation to its other Systemic Features

Maria A. Breiter
Lomonosov Moscow State University / Stanford University
E-mail: polikarp@logos.msu.su

Topical paper

AREA: Quantitatives lexicology.

Summary:
Length of lexemes in Chinese is considered as systemically related to other its important features - frequency, polysemy, part-of-speech and stylistic categories.

1. According to the model of language structure developmentsuggested by A.A. Polikarpov in [Polikarpov, 1979] averagelength of units in sign inventory and its distributional parameters depend, first of all, on the size of inventory. In its turn, the size of inventory depends on general typological shape of a language. More analytic posess more restricted inventories of some basic units. In its turn, degree of the process of analytization depends on the degree and time of the language spread among non-native speakers.

On the contrary, while stable existence of some speech community for a long time without noticeable changes in its members, language system becomes more synthetic.

Tendency for typological change in that or another direction depends on different correlations in different conditions (language spread or language stable functioning) between two kinds of economy - syntagmatic and paradigmatic. In conditions of language spread there is strengthened the relevance for communicative survival of paradigmatic economy. On the contrary in the opposite kind of conditions this kind of economy is relatively less relevant.

In addition we should say that according to the model of language structure dynamics, proposed by R. Koehler [Koehler, 1986] word length as a lexical parameter depends on some important requirements including "minimization of production effort" and "security of transmission". Change of the equilibrium between them in some specific conditions can easily also explaine some regular change of standard length feature of words in some language. But for standard communicative conditions these requirement should be stable for ages. Mainly, the mechanism of interplay between two - "minimization" and "extension" forces works for words of different frequency of use. For the general economy, according to G.K. Zipp it is better to be more waistfüll while producing rare words, but more economical - while producing more usable ones.It leads to well-known dependence of length of units on their frequency of use. But, after all, the whole frequency profile for inventory dependson its size, on language type,

2. As far as Chinese is concerned we can easily observe above-mentionedtendencies.

First of all according to the extremely restricted number of basic units, syllabomorphemes (about 300 unforced and combinational degree are extremely high).

The length Chinese lexical unit is so-called "zi" which usuallycoincides with a morpheme and is represented by one hieroglyph in writing. Being primary Chinese lexical units all the "zi" belong to the conciousness of a language user as distinct from majority ofpolysyllabic units which can be constructed in the process of speaking. This fact is reason of arguing between sinologistswhether Chinese polysyllabic constructions can be considered as lexical units and, in general, are there "words" in Chinese and what are the criteria which can be used to distinguishcompound from word combiations. We distinguish (1) words and (2) lexemesas (1) lexical units and (2) unmarked ones. It means thatthere can be lexemes in some languages without presentce of wordsin them. There can be also such languages which mark some lexemiccombinations of morphemes by some specific grammar affixes, converting these lexemes into words.(On this point see works by V.M. Pavlov [1985]).

"Zi" constituted Chinese on early stages of the language developmentbut since the spheres of life of the language users were broadeningthe need for new nominations was becoming stronger. At the same timethe limited possibilities of phonemic changes and variationsinfluenced on the fact that two main ways to increase the number of meanings turned out to be polysemy (increasing of number of meaningsposessed by each particular language sign) and active combinations of "zi" which gave the possibility of creating new lexemes with new meanings on the basis of already existing language units.

Each syllabomorpheme in Chinese possesses its own clear meanimg,and that is why the meaning of new complex lexemes was in general the result of interaction between the meanings of the constituting elements.Majority (about 83%) of the complex units of the contemporary Chinesecan be easily explained etimologically (e.g. "tose" = "camel + colour" = "yellow-brown colour, colour of a camel", "ribao" = "day + paper" = = everyday paper, etc.). There have been marked out and studied by many linguists standard models of word-formation in Chinese like "generalization" (e.g. "wucai" = "five colour" = "many coloured" and not "five-coloured") (Hi Shida, 1985) or groups consisting of verb and an object with a verbal meaning like "zhidao" = = "know + truth" = "know" or "change" = "sing + song" = "sing"(Ren Xueliang,1981), etc.

This "transparency" and regularity of structure of complex lexical units is the reason why some sinologists consider these complex units not to be regarded as "words" or "lexemes"(Fan Xio, 1980). In its development the Chinese language structure needed more and more such complexes and more and more of these complexes became idiomatic (e.g. "chizu" - "to be jealous", literally: "to eat vinegar"; "ganlantou" - "to behave as everybody", literally: "to catch up with the waves" (Ma Guofan, 1978).

There is also a specific way of formation of polysyllabic lexical units in Chinese which can be explained as a result of lack of the phonetic means to reduce the coding/decoding effort. These units which can be called "shortened (or "reduced") nominations" (Semenas, 1992, p.30) are formed on the basis of expressions including two or even more complex lexemes, e.g. "jianzhong gongye" ("light and heavy industry") is formed out of "jiang gongye zhong gongye" ("light industry and heavy industry"), or "kejie" ("science and engineering") is the reduction of "kexue jieshu" which has the same meaning. This

tendency towards the increase of idiomaticity by throwing away the syllabomorphemes which are not necessary for understanding the meaning of the whole complex is one of the most widely used in Chinese means to increase the number of meanings without increasing the length of lexical unit.

The tendency towards minimization of production effort leads to the formation of "reduced nominations" which include numerals as main components, e.g. "sanhaoshen", literally: "three + good + students" meaning "students who are good in looking for their health, good in work and good in study", or "sida", literally: "four + big" which include "daming" - "to express opinion widely", "dafang" - "to expound point of view widely", "dabianlun" - "wide discussion" and "dazibao" - "newspaper of big hieroglyphs". This kind of reduced complexes in Chinese is formed according to particular models but in order to know the meaning of such complex it is not enough to know the meaning of each hieroglyph, and most of the investigators agree that this units belong to the vocabulary (Wan Dechun, 1983). There is a tendency towards the increase of the number of such complexes in contemporary Chinese.

3. The model of word life cycle describes the regularities of the processes of acquisition of new meanings and loss of old ones by a typical word. According to A. A.Polikarpov, "each sign being originally introduced into language for designation, as a rule, of a certain single meaning, may realize, exhaust its semantic potential in the course of further use and generation of the meanings meanings from the first and derived ones from it. This also means that each of the emerging subsequent meanings of the word is, on the average, more abstract, i.e. has, on the average, a steadily decreasing number of features (components) and associative links with other images and ever lower propensity for "generating" new meanings compared to the previous meanings" (Polikarpov, 1993, p.55).

The development of each subsequent meaning of a typical word towards increasing "emptiness", abstractness means the increase of possibility of using each of the subsequent meanings referring increasingly broadening area of sences. It means, in its turn, that each subsequent meaning of a typical word is in general more and more frequently used. And according to the requirement of the minimization of production effort the more frequently each of the meanings of the word is used (and therefore the word as a whole) the shorter the word has to be. This tendency toward increase of abstractness - increase of frequency - decrease of length needs some quantitative study.

4. The dependencies between the parameter of word length and some other systemic parameters were studied in the present research on the basis of 3000 Chinese hyperlexemes. The term "hyperlexeme" as it is used in this paper was proposed by L. Sacharnij (Sacharnij, 1974) and developed in (Boroda, Polikarpov, 1984; Polikarpov, Karimova, 1989). Hyperlexeme in Chinese is a central unit of lexical system which consists of several lexeme unites on the basis of the same set of lexico-semantic components in, at least, one of the meanings of each lexeme and the same phonetic (or graphic) representation (Breiter, 1992; 1993-I, 1994). The example of Chinese hyperlexeme is "ai" which includes so-called "verb" ("to grieve, to mourne"), "noun" ("sorrow, grief") and "adjective" ("sorrowfull, mournfull") which has polysemy 1, or "hua" which unites meanings and syntactic functions of noun, verb, adverb and adjective and includes meanings "beautiful, bright, shining, flower, colour, fame, to blow, etc." and has polysemy 12 (Breiter, 1993-II).

The sources of hyperlexemes for investigation have been taken from "Frequency Dictionary of Modern Chinese" (Xiandai hanyu pinlu cidian, 1985) and the "Large Chinese-Russian Dictionary" (Bol'shoj kitajsko-russkij slovar', 1994). "Frequency Dictionary of Modern Chinese" includes 31 159 different lexemes which are the result of investigation of texts of the same length which included 1 314 404 running lexemes. The information on polysemy and lexicomorphological classes (which can be compared to parts of speech in European languages) was obtained with the help of "Large Chinese-Russian Dictionary" which gives a great variety of lexical and some other characteristics obtained from the most representative Chinese dictionaries. The sample included 3000 Chinese hyperlexemes: 1000 the most frequent ones (170<F (frequency)<73 835); 1000 hyperlexemes (0<F<2). It can be considered representative because in Chinese there is about 400 untoned and 1300 toned syllables, and 391 untoned and 1021 toned of these syllables turned out to be included into this group. This sample is representative not only quantitatively but structurally because there are represented hyperlexemes of different frequency.

The distribution of hyperlexemes on their length was studied separately for each frequency group and each lexico-grammatical class.As it is shown on table 1, there is a tendency for hyperlexemes of each lexico-grammatical class that more frequent hyperlexemes are shorter in general than the less frequent ones. As far as different lexico-grammatical classes are concerned it can be concluded that the average length of hyperlexeme decreases in direction from the class of nouns (which is in general the more concrete in the meanings and in categorial semantics than the other classes) through the classes of verbs and adjective (which are in general more abstract) to numerals and syntactic words (which are the most abstract classes) (Breiter, 1993-II).This tendency corresponds to the statements of the theory of word life cycle. The more abstract is a meaning of a typical word, the more objects, processes and features it can cover and therefore more frequently it can be used. The more frequently a typical word is used the shorter it is.

The exact form of the interdependence between the length and the frequency of lexical units was studied in: (Koehler, Zoernig, Brinkmoeller, 1990). This dependency which can presumably be considered as universal has an effect on Chinese structure in particular.

The statistical evaluation of the average length for three frequency groups under investigation included calculation of square deviation and confidence intervals for each group: L = 1,55 +/- 0,1063; medium frequency: L = 1,39 +/-0,0387; high frequency: L = 1,23 +/- 0,1254. The intervals do not intersect each other what means that the differences between these frequency groups can be concidered as significant.

As far as the quantitative structure of Chinese hyperlexemes is concerned, the ivestigation pointed out that the regular length of Chinese hyperlexeme is one or two syllabomorphemes. Three- and more syllabomorphemic units are rare and usually it is hard to distinguish between these hyperlexemes and combinations of several hyperlexemes. Three-syllabomorphemic hyperlexemes can be formed by reduction of four-syllabomorphemic combinations (e.g. "daxuesheng" ("student") is a result of reduction composition "daxue xuesheng" ("university student")) or on the basis of a combination of two or three separate hyperlexemes (e.g. "bandaoti" ("semi-conductor"), literrally: "half + lead + heat") or include reduplicated lexemes (e.g. "bailiangliang" (light white"), literally: "light + white + white")). The status of some of this combinations which are considered in the dictionary of Chinese

Table 1.

DISTRIBUTION OF HYPERLEXEMES BELONGING TO DIFFERENT FREQUENCY GROUPS AND LEXICO-GRAMMATICAL CLASSES ON THEIR LENGTH

"part of speech" (for 3 groups) \ length (in syllables)		1		2		3		4		Sum	average length	s*
		absolute	relative**	absolute	relative	absolute	relative	absolute	relative			
"Noun"	1***	102	0,102	364	0,364	0	0,000	13	0,013	479	1,82	139,6
	2***	244	0,244	200	0,200	3	0,003	2	0,002	449	1,47	155,2
	3***	206	0,206	63	0,063	1	0,001	0	0,000	270	1,24	134,4
	Sum	552	0,184	627	0,209	4	0,001	15	0,005	1198	1,57	403,0
"Verb"	1	140	0,140	85	0,085	2	0,002	6	0,006	233	1,45	71,3
	2	180	0,180	66	0,066	0	0,000	0	0,000	246	1,27	47,4
	3	193	0,193	28	0,028	0	0,000	0	0,000	211	1,18	96,9
	Sum	513	0,171	179	0,060	2	0,001	6	0,002	690	1,31	265,5
"Adjective"	1	133	0,133	108	0,108	1	0,001	8	0,008	250	1,53	33,4
	2	197	0,197	25	0,025	1	0,001	1	0,001	227	1,11	45,3
	3	269	0,269	64	0,064	0	0,000	0	0,000	333	1,19	72,8
	Sum	599	0,200	197	0,066	2	0,001	9	0,003	812	1,27	157,4
"Numeral"	1	7	0,007	6	0,006	0	0,000	0	0,000	13	1,46	0,7
	2	8	0,008	0	0,000	0	0,000	0	0,000	13	1,38	10,6
	3	63	0,063	8	0,008	0	0,000	0	0,000	71	1,11	18,4
	Sum	78	0,026	19	0,006	0	0,000	0	0,000	97	1,20	28,3
"Syntactic Word"	1	20	0,020	5	0,005	0	0,000	0	0,000	25	1,20	4,2
	2	58	0,058	7	0,007	0	0,000	0	0,000	65	1,11	14,9
	3	101	0,101	12	0,012	0	0,000	0	0,000	113	1,11	81,3
	Sum	179	0,060	24	0,008	0	0,000	0	0,000	203	1,12	78,9
Total	1	402	0,402	568	0,568	3	0,003	27	0,027	1000	1,55	216,8
	2	687	0,687	305	0,305	4	0,004	4	0,004	1000	1,36	365,3
	3	832	0,832	167	0,167	1	0,001	0	0,000	1000	1,17	448,5
	Sum	1921	0,640	1040	0,347	8	0,003	31	0,110	3000	1,35	917,4

* *s - square deviation*

** *totality of hyperlexemes of each frequency group is regarded as 100%*

*** *1, 2, 3 - frequency groups of hyperlexemes*

1 - low frequency

2 - medium frequency

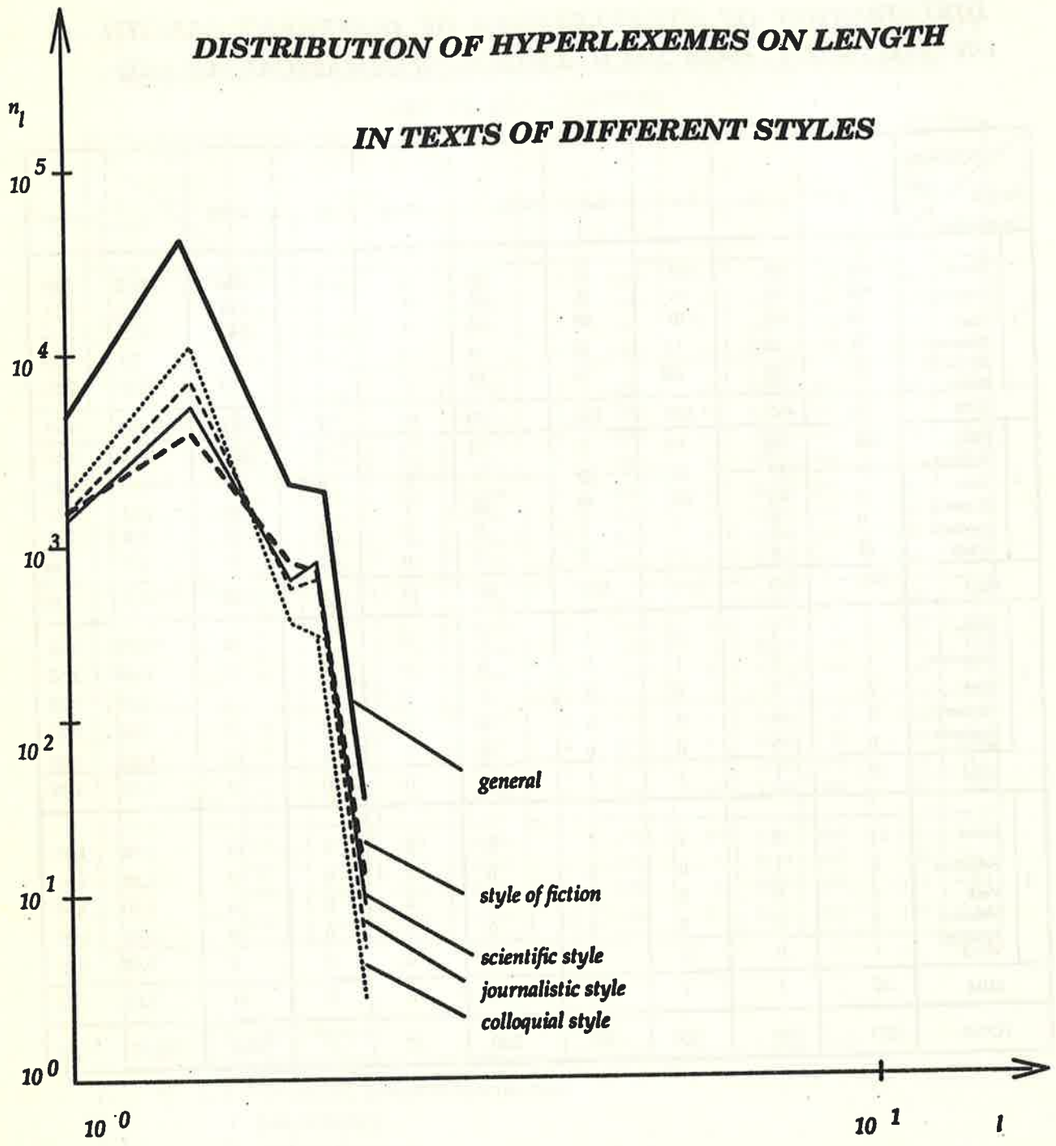
3 - high frequency

Table 2.

DISTRIBUTION OF HYPERLEXEMES OF DIFFERENT LENGTH ON POLYSEMY (FOR EACH LEXICO-GRAMMATICAL CLASS)

length (in syllables) \ polysemy (P)		1	2	3-4	5-8	9-16	17-32	>32	SUM	s	average P
1	Noun	448	168	106	41	28	4	1	796	159,2	3,68
	Adjective	76	97	74	35	52	4	1	339	37,0	6,11
	Verb	148	160	110	66	54	4	0	542	64,5	5,12
	Numeral	26	12	11	10	4	2	0	45	8,5	4,27
	Syntactic Word	93	28	26	14	16	2	0	179	31,6	2,56
SUM		791	465	327	166	154	16	2	1921	280,8	4,23
2	Noun	269	146	77	71	22	2	0	587	96,5	2,44
	Adjective	34	37	22	22	14	2	0	131	15,9	4,48
	Verb	74	43	60	48	28	7	0	260	27,1	3,67
	Numeral	9	6	7	3	0	0	0	25	3,8	2,65
	Syntactic Word	13	8	5	9	2	0	0	37	5,0	1,96
SUM		399	240	171	153	66	11	0	1040	141,1	1,91
3	Noun	2	1	0	1	0	0	0	4	0,79	2,25
	Adjective	1	0	1	0	0	0	0	2	0,49	2,50
	Verb	2	0	0	0	0	0	0	2	0,76	1,00
	Numeral	0	0	0	0	0	0	0	0	0,00	0,00
	Syntactic Word	0	0	0	0	0	0	0	0	0,00	0,00
SUM		5	1	1	1	0	0	0	8	1,77	1,88
4	Noun	11	2	1	1	0	0	0	15	3,98	1,53
	Adjective	8	1	0	0	0	0	0	9	2,88	1,11
	Verb	6	0	0	0	0	0	0	6	3,10	1,00
	Numeral	0	0	0	0	0	0	0	0	0,00	0,00
	Syntactic Word	1	0	0	0	0	0	0	0	0,45	1,00
SUM		26	3	1	1	0	0	0	31	9,57	1,35
TOTAL		1221	709	500	321	220	27	2	3000	322,87	3,06

Picture 1.



l - length of hyperlexemes

n_l - quantity of hyperlexemes possessing length equal to l

as indivisible lexical units is disputable (e.g. "caoluse", literally: "grass + green + colour"; and "cuiluse", literally: "emerald green colour" according to their semantic characteristics should be probably considered as combinations of lexical units). This statement concerns such terminological units as "daiyuejin" ("great advance") which are frequently used by mass-media and therefore possess high frequency but according to their structure probably can not be regarded as lexical units.

5. The distribution of hyperlexemes of different length on polysemy (for each lexico-grammatical class separately) (Table 2) shows that the polysemy of hyperlexeme of low length is higher in general than the polysemy of hyperlexemes of higher length for each lexico-grammatical class. It has been also shown that hyperlexemes of more abstract lexico-grammatical classes are lower in length than the more concrete ones.

6. The average length of the lexical units differs for different styles (fiction, journalism, scientific style and colloquial style) used to compile the Frequency Dictionary of Chinese is 1,9680. The highest index is one of the style of fiction (2,0367), then goes journalistic style (1,9311), and, finally, colloquial style (1,8303) (picture 1). This tendency is natural because the style of fiction, in general, is the most complicated what results in use of a great number of words with low frequency which are usually longer than more frequent ones. The colloquial style is aimed at successful communication and uses the most understandable words which are more frequent and therefore shorter in comparison to those used in other styles.

7. The complete analysis of length of lexical units included into "Frequency Dictionary of Modern Chinese" (Frequency Dictionary..., p. 1489) has shown that its vocabulary consists of 31 159 lexical units and includes 3 751 (12,0%) monosyllabic words, 22 941 (73,7%) - two-syllabic, 2 734 (7,6%) - three-syllabic, 2 010 (6,4%) - four-syllabic, 83 (2,0%) - five-syllabic ones. This tendency towards "two-syllability" of lexical units in Chinese vocabulary can be compared to the tendency of lexical units length distribution in text.

According to the dictionary, units of the text are mostly monosyllabic (64,3% in comparison to 34,3% two-syllabic). The reason of this difference between length of lexical units in vocabulary and in text is obvious: more frequent lexical units (which are therefore more frequently used in text) are usually ones of the lower length as it was stated above.

References

1. Breiter, M.A. (1992) Rol giperleksemnykh gruppirovok slov v realizatsii kommunikativnoi napravlennosti vyskazyvaniya na materiale kitaiskogo jazyka (The Role of Hyperlexemic Groups of Words in Realization of the Communicative Orientation of an Utterance in Chinese) // Problemy semantiki i pragmatiki (Problems of Semantics and Pragmatics). - Moscow: Institute of Linguistics of Russian Academy of Sciences. - P. 105-106.
2. Breiter, M.A. (1993-I) O sistemnykh kharakteristikakh giperleksem v kitaiskom jazyke (On Systemic Characteristics of Hyperlexemes in Chinese) // XXIV Conference "Society and State in China", v. 2. - Moscow: Institute of Oriental Studies of Russian Academy of Science. - P. 159-161
3 Breiter, M.A. (1993-II) Kvantitativnij analiz nekotorykh sistemnykh parametrov leksiki kitajskogo jazyka (Quantitative Analysis of Some Systemic Parameters of Chinese Vocabulary) // Ph. D. Thesis. - Moscow.
4. Breiter, M.A. (1994) Osobennosti vydelenija i struktury giperleksemy v kitajskom jazyke (Features of Hyperlexemic Struture in Chinese) // Kvantitativnaja lingvistica i avtomaticheskij analiz tekstov (Quantitative Linguistics and Automated Text Analysis). - Moscow: Moscow University Press. (In Press).
5. Fan Xiao (1980). Guanyu Jiegou he duanyu wenti (The Problem of a Complex Word and Word-Combination) // Zhongguo yuwen, N 1. - Beijing.
6. Ge Benyi (1985) Xiandai hanyu (Modern chinese). - Beijing.
7. Koehler, R. (1986) Zur linguistichen Synergetic: Struktur und Dynamik der Lexik. - Bochum: Brockmeyer.
8. Koehler, R. (1992) Self-Regulation and Self-Organization in Language // What is the language Synergetics? - Oulu: Univ. of Oulu Printing Center. - P. 17-18.
9. Koehler, R., Zoernig, P., Brinkmoeller, R. (1990) Different equation models for the oscillation of the word length as a function of the frequency // Quantitative Linguistics, vol. 54. Glottometrica 12. - Bochum: Brockmeyer.
10. Large Chinese-Russian Dictionary (1984). Ed. I.M.Oshanin. - Moscow: Nauka. Ma Guofan (1978) Chenyu (Idioms). - Beijing.
11. Polikarpov, A.A. (1993) A Model of the Word Life Cycle // Contributions to Quantitative Linguistics. - Dordrecht, Boston, London: Kluwer Academic Publishers. - Pp. 53-63.
12. Ren Xueliang (1981) Hanyu Zaocifa (Word-Formation in Chinese). - Beijing.
13. Semenas, A.L. (1992) Leksikologija sovremennogo kitajskogo jazyka (Lexicology of Modern Chinese). - Moscow: Nauka.
14. Xiandai hanyu pinlu cidian (1986) (Frequency Dictionary of Modern Chinese). - Beijing.
15. Wan Dechun (1983) Cihui yanjiu (Study of Lexicology). - Jing`an.
16. Zipf, G.K. (1949) Human Behavior and the Principle of Least Effort. - Massachusettts: Reading.

Длина китайского слова в соотнесении с другими системными характеристиками

Брейтер М.А.

Резюме:
Исследуется параметр длины китайского слова во взаимодействии с другими параметрами, включая лексическую полисемию и частоту употребления. Длина исследуется дифференцированно для текстов четырех разных стилей. Сопоставляются характеристики длин лексических единиц и текста.

Some Quantitative Data on the Mashine-Rreadable Version of KSSJ

Vladimir Benko
Comenius University
Faculty of Education
Computational Linguistics Laboratory
SLOVAKIA, SK-81334 Bratislava, Moskovska 3,
Fax: +42-73-62-505

Eduard Kostolansky,
Comenius University,
Faculty of Education
Computational Linguistics Laboratory
SLOVAKIA, SK-81334 Bratislava, Moskovska 3,
Fax: +42-73-62-505

Topical paper

AREA: Computational lexicography

Summary:
KSSJ (Kratky slovnik slovenskeho jazyka (2-nd ed., 1989) - The Concise Dictionary of the Slovak Language) is a one-volume dictionary of present-day Slovak (cca. 65,000 entries). Its MRD version, containing the complete 2-nd Edition information, has become a basis for the building up of the Slovak Lexical Database that is now being enriched by other linguistic data. The paper contains the basic characteristics of the KSSJ and some statistical data obtained by special-purpose analysis.

1. Introduction

Our project on computer-aided Slovak language processing has been started by building up a Slovak computer lexicon. Arguments in favour of such a decision come mainly from our conviction that computational linguistics research has to be tied up strongly with natural language processing (NLP) applications, and that any non-trivial NLP application requires more than a laboratory-size lexicon.

Computer lexicon development favours an iterative methodology [Kostolanski, 1991]. The basic idea here is that a usable product is to be produced in every iteration cycle. This will then become input into the next cycle, where it will be further improved. Some cycles may contribute to the improvement of the technology itself. In our case, an iteration cycle contributes to the improvement of both the lexicon and the NLP application by their mutual interaction.

2. Main characteristics of the MRD-version of KSSJ

The MRD-KSSJ has been derived from the "typesetting tape" - a set of floppy disks in our case - which had been acquired from the publisher. Processing of the tape followed the traditional step-by-step methodology [Byrd et al., 1987] that can be roughly described as consisting of three conceptually independent operations: conversion, filtration and normalization. Some technical details of the process can be found in [Benko,1991]. At present, the MRD-KSSJ contains the complete information from the 2nd printed Edition. The micro- and macro-structure of the dictionary is marked up by a set of (one-character) tags, derived from the typesetting commands, and headword identifiers that have been generated by a specialized program.
Entries in the MRD-KSSJ look like this:

```
!b054c14
"bunda"   -y bund |z|
   {1} kratky sport. kabat, vetrovka (s kozusinou):
 'lyziarska, kozena b.'
   {2} |zastarav.| kozuch: 'zababusit sa do b-y'
   {3} |expr.| husta srst (psa), vlna (ovce) ap.;

!b054c14a
"-icka"      -y -ciek   |z zdrob.|
!b054c15
"bundas" -a |m hovor.| pes s hustou dlhou srstou
```

```
|  |||  |
|  |||  |-------> definition
|  ||  |--------> stylistic label
|  |  |---------> part-of-speech label
|  |------------> morphological information
|---------------> headword
```

KSSJ entries have the traditional explanatory dictionary structure, similar to, e.g. [Oxford Advanced Learner's Dictionary, 1989] or [Webster's 2-nd New Riverside University Dictionary, 1984]. The main headword is followed by morphological information (inflected forms) and a part-of-speech label. Some entries may contain information about pronunciation and/or etymology. Regular derivates, some types of homonyms and aspect or reflexiveness-bounded groups of verbs are nested into one dictionary paragraph and appear usually in abbreviated form. The definition part of the entry contains a sense description and/or a series of synonyms/antonyms. Other elements appearing in the KSSJ entries are: example phrases, stylistic and normative labels, usage notes, lexicalized collocations, phraseology, references, etc. The linguistic information categories are marked up by means of four different typefaces, punctuation and special graphic symbols. (Some entry elements are indicated in the second entry of above picture).

The first step to be done in the MDR-KSSJ development was the disambiguation of graphic representation of the individual information categories. This has been achieved after several iterative steps and the text is now considered to be coherent. In the next step, a program to generate full forms of the abbreviated nested headwords has been written. Similarly, the full forms of headwords in the example phrases have been produced (partially manually). After this iteration the reference version of the MRD-KSSJ was reached that became basis for a lexical database and several linguistic applications. The size of the reference version has increased by 15 % to some 7 MB.

3. Some results.

The analysis, the results of which are described below, has been performed by a set of software tools, including the WordCruncher dBase IV and a series of programs written in C and LEX.

There are good reasons to consider the KSSJ a proper representative of the Slovak vocabulary (except for proper and other names). The distribution of vocabulary according to the part-of-speech labels in dictionary entries is given in the following table:

Nouns	22535	38.0 %	Prepositions	200	0.3 %
Verbs	16342	27.6 %	Conjunctions	114	0.2 %
Adjectives	11960	20.2 %			
Adverbs	4003	6.7 %	References	2440	4.1 %
Pronouns	285	0.5 %	Abbr. + Symbols	360	0.6 %
Interjections	283	0.5 %	Other	206	0.3 %
Numerals	281	0.5 %			
Particles	261	0.5 %	Total entries	59270	100 %

Slovak nouns may have three genders: masculine, feminine and neuter. The verbs may have imperfective, perfective or iterative aspects. The quantities of the individual genders and aspects in the KSSJ entries are as follows:

Masculine	9082	40.3 %	Imperfective	7108	43.5 %
Feminine	10516	46.7 %	Perfective	8113	49.6 %
Neuter	2257	10.0 %	Iterative	252	1.5 %
Other	680	3.0 %	Other	869	5.3 %
Total nouns	22535	100 %	Total verbs	16342	100 %

4. Miscellaneous

The MRD-KSSJ contains 36,032 dictionary paragraphs; 21,000 (sic!) of them contain nested entries. The number of originally abbreviated headwords is 15,778 (26.6 %). The most homonymous forms has the word "zapierat" (1--3 = imperf. from "zapriet" to deny/to push/to close and lock, 4 = imperf. from "zaprat" to wash in(?)). The largest number of the meanings have the following headwords:

prejst (to pass)	24
prist (to come)	20
ist (to go)	16
vytiahnut (to pull)	16

The most frequently used stylistic labels are: expr. (expressive) - 5323, hovor. (colloquial) 2828, kniz. (bookish) 1514 and pejor. (pejorative) 772.

The longest KSSJ headword is "dialektickomaterialisticky" (adjective, dialectical-materialistic), the longest non-composite word is "skomercionalizovat" (perfecive verb, to commercialize).

5. Conclusion and the next work

The development of the MRD-KSSJ has been a tedious, loitered, and from time to time also a "black" work. This work, however, was a prerequisite for further research in the area of Slovak computer-aided lexicology and lexicography. Every piece of information that has been normalized and/or added to the MRD-KSSJ will be usable in the next linguistic projects. For example, the morphological information from the MRD-KSSJ has been used already to build the Slovak morphological data base. Our other on-going project will use the results of the analysis of data links representing the polysemic and hierarchical relations that exist among individual elements of the dictionary entries. Without a doubt, the detailed analyses of the lexicon promise an interesting work.

Квантитативные данные по машиночитаемой версии КССЯ

Бенко В., Костолански Э.

Резюме:
КССЯ - "Краткий словарь словацкого языка" (2-е изд., 1989) является однотомным словарем современного словацкого языка (около 65 тыс. статей). Его машиночитаемая версия, содержащая полную информацию по 2-му изданию, стала основой для построения базы данных по словацкой лексике, пополняемой другими лингвистическими данными. Доклад содержит базовые характеристики КССЯ и некоторые статистические данные, полученные в ходе специального анализа.

Опыт автоматизации исследований русского силлабо-тонического стихосложения

И.Е.Воронина, А.А.Кретов, А.Суворов
Воронежский университет
Россия, 394693, Воронеж, Университетская пл. 1
E-mail: fna@amm.vucnit.voronezh.su

Доклад

ТЕМАТИЧЕСКАЯ ОБЛАСТЬ:Квантитативные исследования русского стихосложения

Резюме:
Описывается программа автоматического анализа русского стихосложения.

Главным достоинством компьтеров является их способность выполнять трудоемкую и нетворческую интеллектуальную работу. В исследованиях по стиховедению именно такая работа - по первичной обработке и накоплению статистически достоверного материала занимает наибольшее время. Освободить исследователя от рутинной работы и сэкономить его время призвана программа AVERS (Automatic Analysis of Versification).

Данная программа обрабатывает текстовые файлы, содержащие русские стихотворные тексты с проставленными ударениями, и ориентирована на силлабо-тоническую систему стихосложения.

Программа анализирует размер каждой строки (ямб, хорей, дактиль, анапест, амфибрахий), рифму по месту ударения (мужская, женская, дактилическая, гипердактилическая), наличие спондеев и пиррихиев. Кроме того, она дает статистику по каждому стихотворному тексту: количество размеров, число строк и процентное соотношение по каждому из них; количество пиррихиев и спондеев в стихотворном тексте; количество каждого типа рифм (по месту ударения) и расстояние (в строках) между рифмами.

Описанные в литературе алгоритмы определения размера и рифмующихся строк показали невысокую разрешающую способность, что потребовало разработки более эффективных алгоритмов.

По желанию пользователя на экран выдаются строки, содержащие тот или иной параметр анализируемого стихитворения. Особый файл содержит вокалическую структуру стихотворения, разбитую на стопы, что дает богатый материал для исследования ассонансов.

В настоящее время программа является исследовательской и может применяться при проведении широкого круга научно-исследовательских работ: от курсовых до диссертационных.

На базе имеющейся версии программы предполагается создание обучающей программы учащихся средней и высшей школы, которая будет знакомить с основными понятиями русского силлабо-тонического стихосложения и тренировать в практическом пользовании ими.

An Attempt for Automatization of Russian Syllabotonic Versification Study

Voronina I.E., Kretov A.A., Suvorov A.

Summary:
It is presented AVERS - a program packet for automatic analysis of Russian versification.

The Problem of Measuring Linguostatistic Pecularities of Author's Speech in Fiction Texts

G.V.Ermolenko
Grodnensky Pedagogical Institute
Grodno, Bielorussia

Topical paper

AREA: Quantitative stylistics

Summary:
It is suggested two linguo-statistic indices wich allow to measure picularities of author's style. This approach is used to test hypothesis of N. Medvedeva about existing of two authors in text of novel "Тихий Дон" by M.A. Sholohov

1. Measuring in linguistics brings definite precision and reliability in the results of research, promotes approaching linguistics to exact sciences, advantageously distinguishes such descriptions of language phenomena from those, constructed upon subjective impressions of an investigator.

2. It is undertaken measuring peculiarities of author's speech in fiction prose with the help of statistical indeces:

1) index of using prepositional and conjuctional connections

$$R = \frac{\sum_{i=1}^{m} (P_i - C_i)}{m},$$

where P_i is the number of prepositions, C_i - the number of conjuctions in m - portions containing 100 running words;

2) index of prevalence of zero occurencies of Russian conjunction in the structure of homogeneous verbal predicates

$$Q = \frac{\sum_{i=1}^{m} И_i}{n},$$

where И - the number of cases of using the conjunction И before the last part in range of homogeneous predicates, which is situated directly after a comma; n - the number of sentences.

3. It is undertaken comparison of dimensions R and Q, which measure the peculiarities of aspect of expression on syntactical level, using the criterion of comparison of mean arithmeticaldimensions of the variational series. The criterion looks as follows:

$$\& = \frac{\bar{X}1 - \bar{X}2}{\sqrt{\frac{@2}{X1} + \frac{@2}{X2}}},$$

where $\bar{X}1$ and $\bar{X}2$ - mean arithmetical dimensions of compared variational series,
@2 @2
-- and -- squares of the
X1 X2
main errors of selection, @ - "sygma", & - "gamma".

4. It is undertaken examination of N. Medvedeva's hypothesis in her book "Стремя "Тихого Дона" (Mysteries of the Novel). - Paris: YMKA - Press, 1974, about presence of two author's origins (author - "co-author") in the text of the novel "Тихий Дон" by M.A. Sholokhov.

Проблема измерения лингво-статистических особенностей авторской речи в литературных текстах

Ермоленко Г.В.

Резюме:
Предлагаются два лингвостатистических индекса, позволяющих измерить особенности стиля автора. Этот подход используется при проверке гипотезы Н. Медведевой о существовании двух авторов текста романа "Тихий Дон" М.А. Шолохова.

From Latin To Modern Romance Languages: Testing Regularities for Words System Features Evolution

Kapitan, Maxim E.,
Moscow State University
Philological Faculty
Dept. of Romance Philology
Russia, 117899, Moscow,
Vorobjovy Gory, MGU, 1 Bldg. of humenetes
Philological Faculty

Topical paper

AREA: Description of language evolution process

Summary:
It is found that the degree of survival of classical Latin words in some major modern Romance languages depends on their age, polysemy and part-of-speech category. Besides, the dependences of values of part of speech category, polysemy, frequency, word-building structure on their age are considered.

1. In this paper the dependence of Latin words survival degree in major modern Romance languages on their age has been considered. Besides, we undertake an attempt to test the dependence of words of different parts-of-speech, polysemy zones, frequency ranks, word-building structure on their age.

Also an interplay of some indicated systemic features of words without age differentiation was investigated.

Some major tendencies for predisposition of Latin words of different categories, having different systemic features to survival and some clear system diachronic transformations have been revealed in the model of evolutional development of lexical system, based on the concept of word life cycle by A.A.Polikarpov [Polikarpov, 1988, 1990, 1991, 1993, 1994].

This model enables to forecast the greater survival chances for words which are relatively more ancient, categorically and semantically more abstract, more polysemic, more frequent, with simpler word-bilding structure.

2. The first thousand of the most frequent words from D.D.Gardner's Frequency Dictionary of Classical Latin Words [Gardner, 1970] was taken for the test of forecasted correlations. The age of Latin words has been determined using data from Ernout-Meillet etymological dictionary [Ernout A., Meillet A., Dictionnaire etymologique de la langue latin. Histoire des mots. 1959]. We used age divisions as follows:

1 - period of all-indoeuropian unity,
2 - period of western-europian unity,
3 - period of Italic unity,
4 - strictly Latin period.

(In some cases 1-3d periods were considered like one period, opposed to the 4th, because relationships between some characteristics, marked by these the most ancient periods, in these cases are not obvious.)

Polysemy was controlled in P.G.W.Glayr's explanatory dictionary of classical Latin words [Oxford Latin Dictionary, 1968-1982]. Survival rate was defined according to W.Meyer-Lubke's etymological dictionary [Meyer-Lubke,1935].

The following results were obtained:

Table 1

Correlation between age and survival for words of all parts of speech together

p	languages											
	Latin		Rumanian		Italian		French		Spanish		Portuguese	
	abs.	%	abs.	%	abs.	%	abs.	%	abs.	%	abs.	%
e	238	100	107	44,96	150	63,03	132	55,46	144	60,5	151	63,45
r 2	78	100	31	39,74	46	58,97	40	51,28	43	55,1	44	56,41
i 3	15	100	4	26,67	8	53,33	6	40,0	7	46,6	7	46,67
o 4	650	100	130	20,0	254	39,08	205	31,54	241	37,0	240	36,92
d	—		—		—		—		—		—	
ALL	981	100	372	37,9	458	46,7	383	39,0	435	44,3	442	45,1

It is clearly seen that per cent of survival of more ancient words is higher for all the five main Romance languages.

3. If we consider the dependence of integral degree of survival of Latin words in modern Romance languages (see Table 2) we can see words with maximum survival (preserved in five major Romance languages) to be more among the most ancient words, but those unpreserved in no one of Romance languages are more among the younger words.

Table 2.

Dependence of integral degree of survival on the age

p	Latin survival									unpreserved	
	5	4	3	2	1	sm	dial	all	%		%
e 1-3	348	106	63	21	27	20	7	8	252	72,4	96 27,6
r i 4 o d	648	75	98	38	47	47	19	25	359	55,4	289 44,6

(5,4 etc. - preserved in 5, 4 etc. major Romance languages, sm - preserved in small Romance languages, dial.- preserved in dialects).

4. ther systemic characteristics of Latin words depend on the age as well:

4.1. orrelation between age and word-building structure

Table 3.

periods	root words	derivatives and compounds	in all
1 - 3	347	2	349
4	150	494	644
	—	—	—
ALL	497	496	993

In the table 3 it is seen that number of root words to be much more among ancient than among relatively young Latin words of this frequency group.

4.2. orrelation between age and belonging to some part of speech

Table 4.

	periods 1 - 3		period 4		in all
	abs.	%	abs.	%	
N	131	36,9	224	63,1	355
V	111	34,37	212	65,63	323
ADJ	60	33,71	118	66,29	178
ADV	17	19,5	70	80,5	87
NUM	3	60,0	2	40,0	5
PRON	3	50,0	3	50,0	6
PREP	9	69,23	4	30,77	13
CONJ	12	46,15	14	53,85	26
INTERJ	2	66,6	1	33,3	3

In the table 4 one can see more ancient words to predominate among numerals and prepositions, and younger words -among autonomous parts of speech. Pronouns and conjunctions occupy an intermediate position.

4.3. orrelation between age and polysemy

Table 5.

p	1 period	2 period	3 period	1-3 period	%	4 period	%	in all
o 1	10	2	2	14	46,7	16	53,3	30
l 2	9	5	1	15	28,3	38	71,7	53
y 3-4	48	18	4	70	32,9	144	67,1	213
s 5-8	73	24	4	101	29,0	247	71,0	348
e. 9-16	78	30	3	111	37,6	184	62,4	295
m 17-46	33	4	1	38	67,9	18	32,1	56
y	234	83	15	349		647		995

4.4. orrelation between age and frequency

Table 6.

frequency ranks	1-3d periods		4th period	
	abs.	%	abs.	%
1-64	38	60,3	25	39,7
65-128	40	62,5	24	37,5
129-256	57	44,9	70	55,1
257-512	84	32,8	172	67,2
513-1000	130	26,7	357	73,3

In the table 6 you can see that with the drop of frequency (with the exception a zone of the most frequent words) the share of ancient words decreases, and the share of young words increases.

5. The data about correlation between survival of lexemes and their polysemy were obtained.

Table 7.

Dependence of survival of lexicon on polysemy for five major modern Romance languages

p	Lat.	Rum.	%	Ital.	%	French	%	Span	%	Port	%
o 1	30	9	30	10	33,3	8	26,7	9	30	10	33,3
l 2	53	6	11,3	16	30,2	11	20,8	19	35,9	18	34,0
y 3-4	213	41	19,3	79	37,1	59	27,7	63	29,6	64	30,1
s 5-8	348	89	25,6	151	43,4	127	36,5	144	41,4	149	42,8
e 9-16	295	97	32,9	164	55,6	142	48,1	160	54,2	160	54,2
m 17-46	56	29	51,8	37	66,1	35	62,5	39	69,6	39	69,6
y	995	265		457		382		434		440	

From the table 7 it is seen, that with the rise of polysemy the per cent of survival steadfastly increases (monosemantic words are an exception to the general rule and it is. probably. explained by a peculiarity of the material: only particularly stable monosemantic words are included into the first thousand of the most frequent words).

Table 8

Dependence of an integral degree of survival of lexemes on their polysemy

polys. zone	in Lat.	preserved 5	4	3	2	1	sm.	dial.	all	%
1	30	5	2	2	2	3	0	0	14	46,7
2	53	3	8	4	4	2	1	1	23	43,4
3-4	213	25	25	12	17	11	7	9	106	49,8
5-8	348	64	51	15	30	36	10	11	217	62,4
9-16	295	68	76	18	20	15	6	9	212	71,9
17-46	56	24	7	7	2	1	2	3	46	82,1

(5,4 etc. - preserved in 5, 4 etc. major romance languages, sm. - preserved in small Romance languages, dial.- preserved in dialects).

From the table 8 it is seen as well, that with the rise of polysemy the part of remained words in a more number of mo

dern languages increases and the amount of unpreserved words decreases.

6.The facts about correlation between average polysemy and integral degree of survival were also obtained.

Table 9.

orrelation between average polysemy and integral degree of survival

5	4	3	2	1	unpreserved
10,75	8,52	8,71	7,04	6,93	6,62

In the table 9 you can see, the more preserved words, the more their average polysemy.

7. The data about correlation between survival and belonging to parts of speech were obtained.

Table 10.

orrelation between belonging to parts of speech and survival rate

	Lat	Rum	%	Ital	%	French	%	Span	%	Port	%
N	355	102	28,7	196	55,2	167	47,0	177	49,9	183	51,6
V	323	85	26,3	136	42,1	108	33,4	138	42,7	138	42,7
ADJ	178	49	27,5	84	47,2	70	39,3	77	43,3	79	44,4
ADV	87	11	12,6	19	21.8	15	17,2	19	21,9	18	20,7
NUM	5	5	100	4	80,0	4	80,0	5	100	5	100,0
PRON	6	5	83,3	5	83,3	4	80,0	5	83,3	5	83,3
PREP	13	7	53,9	6	46,2	7	53,9	8	61,5	8	61,5
CONJ	26	6	23,1	7	26,9	6	23,1	5	19,2	5	19,2
INTERJ	3	1	33,3	1	33,3	1	33,3	1	33,3	1	33,3

According to these facts one can see that the syntactic words to be more preserved than the autonomous words (however conjunctions represent the exception to the general rule, but the reasons of this fact must be considered particularly).

8. The obtained results corroborate the model prediction that the most ancient, most polysemantic, most frequent Latin words remain in the greater number of new Romance languages and in the greater proportion. The significant correlation between such systemic parameters of Latin words as an age, polysemy, frequency, belonging to the parts of speech, was also found.

References

1. Ernout A., Meillet A. [1959]. Dictionnaire etymologique de la langue latin. Histoire des mots. 4-ed. - Paris, 1959.

2. Gardner D.D. [1970]. A Frequency Dictionary of Classical Latin Words. - Stanford University, 1970.
3. Meyer-Lubke W. [1935]. Romanisches etymologisches Worterbuch. - Heidelberg, 1935.
4. Oxford Latin Dictionary [1968-1982]. Oxford Latin Dictionary. - Oxford: Clarendon Press, 1968-1982.
5. Polikarpov A.A. [1988]. K Teorii Zhiznennogo Tsikla Leksicheskikh Yedinits (Towards the Theory of Life Cycle of Lexical Units) // Prikladnaya Lingvistika i Avtomaticheskiy Analiz Tekstov. Tez. dokl. nauch. konf. 28.01-30.01.1988 (Applied Linguistics and Automatic Text Analysis. Papers from the Conference 28.01-30.01.1988).- Tartu: Tartu University Press, 1988.
6. Polikarpov A.A. [1990].Leksicheskaya polisemiya v evolyutsionnom aspekte (Lexical Polysemy in Evolutionary Aspect)//Linguistica-1990 (Acta et Commentationes Universitatis Tartuensis, No 911). - Tartu: Tartu University Press, 1990.
7. Polikarpov A.A. [1993].On the Model of Word Life Cycle // Koehler, R., Rieger, B.(eds.) Contributions to Quantitative Linguistics. - Dordrecht: Kluwer, 1993.
8. Polikarpov A.A. [1994]. Zakonomernosti Zhiznennogo Tsikla -Slova i Evolutsii Yazika.Statja 1.Modelirovanie Osnovnych Sistemnych Sootnosheniy. (Regularities of Word-Life Cycle and of Evolution of Language. Article 1. Modelling of the Main Systemic Correlations.// Russkii Filologicheskiy Vestnik (Russian Philological Bulletin), N 1, tom 79, Moskva: Moskovsky Litsey, 1994.

От латыни к современным романским языкам: проверка закономерностей эволюции лексической системы

Капитан М.Е.

Резюме:
Обнаружено, что степень сохранности классической латыни в современных романских языках зависит от их возраста, полисемии и категории части речи. Кроме этого учитываются зависимости значений категории части речи, полисемии, частоты, структуры словообразования от их возраста.

Affinity of Phonetical and Graphic Representations of the Basic Units of Chinese

A.M.Karapetjanc
Lomonosov Moscow State University
Institute of Asian and African Countries
Chair of Chinese Language
E-mail: polikarp@logos.msu.su

Topical paper

AREA: Length Distribution of Phonetical and Graphic Units in Chinese

Summary:
The basic unit of Chinese - the word-sign is represented by a certain syllable and a certain character. The present investigation is found on the basic list (BL) of 4212 word-signs - the sum of three independent lists of common characters, evolved in the PRC either intuitionally or on the basis of currency dictionaries, and the first part of the library of Chinese characters of CCDOS (the Chinese national standard GB 2312-80). BL extensively covers the set of word-signs, used by the native speakers, and combines the characters common for humanities and sciences. The aim of the investigation is the analysis of the correspondance between phonetic and grafic complexity (length) of wordsigns.

The length of a Chinese syllable can be measured in the number of microphonemes - realized structural units of syllable accordingly to the model

"initial (I) + medial (M) + central (C) + terminal (T)",

i.e. in the number of letters by which the syllable can be optimally represented.

The data on the connection between the phonetical satuation of a syllable and its duration in a text supposes 8 ways of measurement of phonetical length, because there are possibilities of: a) counting aspirated and spirant initials as two microphonemes; b) taking into consideration the reduction of C=e in the case of simultaneous realization in a syllable of M and T; c) ascribing of dual duration to resonant terminals "-n" and "-ng". So there are 8 models: SUN, SUEN, SUNN, SUENN, SSUN, SSUEN, SSUNN and SSUENN (these indications are correlated with the syllable "sun").

The criterium of the linguistical consistency of a certain model is the monotonousness of diminishing of the number of characters, corresponding to a certain syllable, as long as the length of the syllable is increased. This phenomena takes place only for the models SUN and SSUN, in the first case juxtapposed are the legths 1-2 and 3-4, in the second - the lengths 1-3 and 4-5. The additional criterium is the monotonousness of increasement of medium number of strokes as far as the pnonetic complexity of a word-sign increases, because more simple units must have greater productivity; this phenomena is obvious only for the SUN-model. This model is also the best from the point of view of the minimality of the coefficient of variation of the general distribution of productivity and the similarity of juxtaposition of pnonetical and graphical lengths.

The graphic comlexity of a word-sign must be measured in number of graphemes. The number of strokes is not suitable because of great variation and the presence of comlex strokes. In the standard typographic set of 7000 characters the characters beginning with a combined stroke have one stroke less. That means that complex stroke is psychologically conceived as a combination of two strokes in spite of the fact, that it is traditionally counted as a single stroke. It is also to be noted that the productivity of a phonetic part of a character is related more regularly to its length in SGC (see below), not in strokes.

The analysis of graphics begins with the evolvement of the set of immediate graphic constituents (IGC) on the basis of formal isolation of phonetic parts (PP) - the common constituents of the characters with similar pronounciation. The characters of BS have 1288 PP. The regular remainders of characters after deduction of PP can be organised into a list of 100 (with zero) determinatives (DT) - the frequent and simple graphic constituents. There are also "exotic" graphic constituents (EGS), which can be parts of not more than three characters.

The analysis of the lists of modifications of PP, EGS and characters with two determinatives show that a character can comprise no more than four graphic constituents, but the characters with 3-4 constituents occupy less than 5% of BS. The characters having only a PP constitute 25.4% of BS, those having a PP and a DT - 67.5%. The number 4 suggests the greatest number of graphemes in a character.

The distribution of characters according to the number of DT and PP shows a linear correspondance between productivity and the number of strokes and allows the grouping of DT according to their productivity. The form of correspondance of productivity of PP and their length in strokes shows that for such parts to have less than four strokes is not typical. At the same time the DT with 3 strokes have the biggest productivity. The presence of a breaking point in the linear correspondance of number of PP and their productivity in the bilogarithmical scale gives the border between ordinary and very productive PP. All these phenomena prove the fact, that there is no fixed border between semantic and phonetic constituents of characters: there are unusual rare DT and very productive PP.

The correspondance between the number of additional strokes in a character and the number of strokes in its DT shows that the number of additional strokes in characters with rare DT is firmly decreasing by the increasement of the length of the DT (the zero DT is not an exeption). This fact proves that among rare DT the percent of DT, consisting of more than one grapheme, is increasing by the increasement of number of strokes in these DT, because the characters have a tendency for equal length in significal units. At the same time the number of additional strokes in characters with frequent DT does not increase by the increasement of their length and remains on the level of number of additional strokes for common DT with three strokes. This draws us to

the conclusion that all frequent DT are graphemes and not less than half of all DT are identical with graphemes.

The application of the notion of grapheme to other graphic constituents supposes a dissection of PP, which is conducted in two stages accordingly to formal procedures. Firstly is evolved a list of 490 united grahic constituents (UGC) - the maximum graphic intersections of PP, representing parts of considerable number of characters. The connection between the ability of UGC to constitute a part of another UGC and the number of its strokes gives the measure of complexity which suppose a dissection of SGS and the measure of simplicity which allows it to be considered as standard graphic constituent (SGC). The additional criterium is the combination of the ability of UGC to be a part of another UGS with its productivity in formation of characters. In this way on the second stage is formed the list of 250 SGC and it becomes obvious that a standart PP comprises two SGC.

The similarity between the grouping of SGC according to their frequency with their grouping according to their positions in characters (initional, medial and terminal) and their usage as DT allows to discern 7 layers of frequent SGC; each of them, whith the exeption of the last one, comprises circa 16 SGC. There was evolved also a list of 55 graphemes (GRF) by the combination of SGC according to principle of additional distribution.

The dinamics of increasement of accumulated frequences of basic graphic constituents of characters in the text, SGC and also of determinatives, formally postulated for each character, shows the reliability of the supposed list of graphemes. For the characters of BS there are only 22 cases of codes omonimy. The discernability of characters in SGC is only slightly inferior to their discernability in strokes, because in most of these cases the stroke codes are also identical.

All abovesaid allows to see in SGC analogies to the microphonemes. The difference between the number of microphonemes (approximately 30) and the number of SGC is not substantial, because 98 SGC (with the layer 7 - 106) cover more than 80% of realisations of SGC in characters of BS, and the number of GRM can ben minimalized. There is also a certain similarity between DT and initials, PP and finals of syllables.

Among two graphical models SGS-model is better from the point of view of juxtoposition of lengths 1-2 and 3-4, but in the GRF-model the preference of SUN-model is obvious for any length of the word-sign. Those two models can be estimated as having equal value, because the differencies are slight and the segmentation into SGC was a result of more formal procedures.

The statistical characteristics of distribution of word-signs according to their phonetical and graphical lengthes (three middles with variety coefficient, central moments, excess and assimetry) allow to speak not only of their similarity (which is formally false, because they are not samples), but of their identity. It is to be noted that certain characteristics place the SUN-distribution between SGC- and GRM-distributions - two "versions" of the same graphic model. All this allows us to say that the phonetical and graphical lengthes of the basical units of Chinese, measured in linguistically consistent units - microphonemes and graphemes, are gomomorphous.

Statstical characteristics of word-signs according to their length

Model	X	V	G	H	M2	M3	M4	E	A
SUN	2.73	27	2.62	2.49	0.53	-0.05	0.76	2.75	-0.12
SGC	2.71	30	2.58	2.43	0.64	0.04	1.20	2.90	0.08
GRF	2.79	29	2.66	2.51	0.65	0.02	1.26	3.00	0.04

Неопределённость фонетических и графических репрезентаций базовых единиц в китайском языке

Карапетянц А.М.

Резюме:
Исследуются закономерности распределения по длине базовых единиц, слогоморфем, в китайском языке фонетической и графической его репрезентациях. Устанавливается соотношение между фонетической и графической сложностью (длиной) однослогов.

CORRELATION BETWEEN MONO- AND MULTI-DIMENSIONAL UNITS WITHIN SUFFIXAL INVENTORY OF MODERN UKRAINIAN LANGUAGE
(To the problem of determination of order parameters in language system)

Eugenia A.Karpilovs'ka,
O.O.Potebnya Institute of Linguistics
at the Academy of Sciences of Ukraine

Ukraine 252001 Kiev-1, Hrushevsky str.,4
O.O.Potebnya Institute of Linguistics at
the Academy of Sciences of Ukraine

Phone: (044) 228-71-82 or (044) 229-02-92
Fax: (044) 228-53-27

Topical paper

AREA: Quantitative analysis of systemic relations in morphemics

Summary:
Substantiation and verification of the scientific hypothesis about correlation between mono- and multi-dimensional units within the morphemic inventory as one of order parameters in language system

Morphemics and word-generating, order parameters in language system

The hypothesis about correlation between mono- and multi-dimensional units within the morphemic inventory as one of order parameters in language system is substantiated and verified on the data of Modern Ukrainian suffixal inventory. Both the mono- and multi-dimension of the form and semantics of suffixes is analysed and 5 groups of units are formed on this basis. Multi-dimensional units, as it is proved, are the results of the economy of the resources of morphemic inventory in the process of word-generating due to the expansion of its internal possibilities and not due to the increase of its rate.

The morphemic level of language system has, as it is known, the inventory of it's base units - morphemes and the rules of their functioning within it's complex units - word morphemic structures. The latter consist of the 1) rules of morphemes' combination in the linear chain (the rules of morphemic syntagmatics) and 2) the rules of their substitution (the rules of morphemic paradigmatics). The finite result of generating of word out of morphemes is submitted, in it's turn, to the laws of word constructing, which function in certain language and determine the choice of inventory units, order of their succession within the word structure, number of morphemes of certain class and the whole number of morphemes in the word, modes of their formal mutual adaptation (morphonological rules). The study of the functional features of morphemes permits to determine the laws of action of the word-generating mechanism in certain language and on this basis to organize the proper morphemic inventory according to the functional load of it's units. This latter makes possible the prognostication of word synthesis out of certain morphemes.

The aim of proposed investigation was the study of the functional features of the Modern Ukrainian suffixal inventory units, because they are the most active in the processes of word-generating. As the data base for this analysis were used 672 suffixal units, determined in the result of computer treatment of 132,000 simple (one-root) words of Modern Ukrainian language and organized in the frequence- combinatory "Dictionary of morphemes of the Modern Ukrainian language" [1]. The computer determined also all enviroments, in which these suffixes occured, i.e. all their left-sided and right-sided partners in these words. This question was the cardinal for our work: do these suffixes always realize within the words in the same form (writing in letters) and with the same semantics (categorial derivational meaning or function) or both their form and semantics are able to change in certain situations. In other words, do in suffixal inventory really represent formally and semantically mono-dimensional units as well as multi-dimensional ones and if it is so, what kind are they, in the result of what processes they appear, what are their place and functional load in this morphemic inventory.

As the analysis of material proved, within suffixal inventory may be marked 5 groups of units: the suffixes 1) mono-dimensional formally and semantically (MONsMONf - 304, or approx.45,2%); 2) mono-dimensional semantically, but multi-dimensional formally (MONsMULTf -163, or approx.24,3%); 3) multi-dimensional semantically, but formally both monoand multi-dimensional (depending on certain realized meaning - MULTsMONfORMULTf - 103, or approx.15,3%); 4) multi-dimensional semantically and formally (MULTsMULTf - 66, or approx. 9,8%); 5) multi-dimensional semantically, but mono-dimensional formally (MULTsMONf - 36, or approx.5,4%).

Let us underline that we understand multi-dimension of suffix form as it's capacity to realize within the word in the set of formal variants (allomorphs). Units, which are semantically multi-dimensional, we call suffixemes, using this term in interpretation by I.I.Kovalyk [2, p.6] as designation of units, which are able to realize in the word in the set of morphs-homonyms (with various

categorial derivational meanings) or submorphs-homographs.

The first, the most numerous, group of units includes grammatical suffixes, flexions of noun case forms, which felt out the modern system of inflexion, unifixes, !.j-9-), elements, which are determined as the remnent product of morphemic analysis of borrowed words and also various connectors between the root and suffixes of full value, which keep, as a rule, "the track" of borrowed unit.

The overwhelming majority of these units ought to suppose that the "defect", "exotic" character of their form or semantics favour the conservation of their mono-dimension, because they, from one side, require another modes of their formal assimilation and, from the another side, bind the development of their semantics. But in this group there are some active derivational elements, which form keeps invariable thanks to their finite position in the chain of word-generating.

To the second group belong suffixes, which semantics remains invariable, but their form can change. They can replace various positions in the chain of word-generating: to be the source units for transformation or to be the result of this one. Besides proper allomorphs, which are the results of action of certain morphonological rules, in our material occure allomorphs, which we call conditional ones. Their appearance in the writing of word in letters is caused by different modes of representation of the same unit in the combinations with various right-sided partners. Suffixes of this group are organized mostly in the pairs (82 out of 163) and more rarely -in the triads (24).

This group of suffixes is narrowly connected with two groups of suffixemes - MULTsMULTf and MULTsMULTfORMONf. Often in the pairs, triades and bunches with the 4,5 and even 6 elements can be organized both semantically mono- and multi-dimensional units. In the process of word-generating can occure contraction of multi-dimensional unit semantics as well as it's expansion, i.e. the transformation of mono-dimensional unit into multi-dimensional one.

There 21 pairs, 13 triades, 4 bunches with 4 elements and on 1 bunch with 5 and 6 elements correspondingly in our material. In a whole 108 units of suffixal inventory are represented in these complexes. Within them the semantics of the source element regulates the concrete transformations of the units' form.

It is not hard to notice that this net crosses with the nets of representatives of other suffixemes (,I,I). The effect of such nets usage for study of generating mechanism of graphemic and phonological structure of the word was convincingly demonstrated by P.Menzerath still in 1954 [3]. As our word proves, it is possible to brigthen up many essential things in the mechanism of morphemic word-generating with their help too.

Within suffixemes it is possible to choose two levels for the contrasting of their representatives: on their status in the word morphemic structure (morphs and submorphs) and on the character of categorial derivational meanings, which are peculiar to the morphs. These latters may demonstrate the homonymy within one part-of-speech meaning, between different part-of-speech meanings and, at least, so-called mixed homonymy, which combine both above-mentioned types.Taking into account various combinations of these modes of contrasting and various types of morphs' homonymy, 7 varieties of suffixemes may be formed. In our material prevail suffixemes with homonymy morphs' meanings within the noun part-of-speech meaning and with the contrasting such homonymic morphs to submorphs-asemantemes. See about it in more details in [4].

The analysis of our data permitted to suppose the existence of correlation between mono- and multi-dimensional units in the Modern Ukrainian suffixal inventory. Multi-dimension of the form as well as multi-dimension of semantics appears due to the economy of means in the process of word-generating, to the frequentative repetition of the same units (from semantic or formal point of view) as possibly more acts of word-constructing. Multi-dimension of form is ensured by the brunched system of morphonological procedures, but mostly they function not deeper than on one step of the chain of generating. On one step it may be from 1 to 5 such formal transformations.

Multi-dimension of semantics is created due to the coincidence of the results of formal transformations with the invariable units. For recognizing of suffixemes language works out special mechanism of diagnostics, which includes various enviroments of such units within the word, their accent features, type of word morphemic structure in a whole.

Thus, units with multi-dimensional form may be organized in various complexes (pairs, triades and bunches with 4-6 elements). The units with multi-dimensional semantics are organized into suffixemes of 7 varieties. On this basis it is possible to determine within inventory: 1) invariable units (group MONsMONf); 2) complexes of semantically mono- dimensional allomorphs (group MONsMULTf); 3) suffixemes with invariable form (groups MULTsMONf and partly MULTs MULTfORMONf) and 4) suffixemes-allomorphs (groups MULTsMULTf and partly MULTsMULTfORMONf). Taking it into account, it is possible also to determine 3 degrees of the ability for the organization of suffixal inventory units, i.e. zero (MONsMONf); the first (on the indication of the form (MONs MULTf) or semantics (MULTsMONf, MULTsMONfORMULTf) change and the second one (on the indication of change both form and semantics of units (MULTsMULTf).

Stepped character of suffixes' organization is caused by their different functional load in the process of word-generating. The active usage of the large part of suffixal inventory in the constructing of word morphemic structure is caused by the tendency to the economy of means thanks to the various procedures of their formal and semantic transformation. The multi-dimensional units have higher functional load and may be considered as the expansion of suffixal inventory possibilities in the process of word-generating due to it's internal resources and not due to the increase of it's rate.

X".", 1962, S.5-26.

REFERENCES

1. Словник морфем української мови. За ред. Н.Ф. Клименко.- Київ,1991.

2. І.І.Ковалик. Питання словотворчої омонімії і синонімії в сфері іменників слов'янських мов. // Питання слов'янознавства. - Львів, 1962, с. 5-26.

3. P.Menzerath. Die Architektonik des deutschen Wortschatzes // Phonetische Studien, 1954, No.3.

4. Е.А. Карпіловська. Формальне варіювання суфіксів у сучасній українській мові. // Мовознавство, 1993, No 5, с. 32-43.

Корреляция между одно- и многомерными единицами в пределах суффиксального инвентаря современного украинского языка.

Карпиловская, А.

Резюме:
Обоснование и проверка научной гипотезы о корреляции (взаимосвязи) между одно- и многомерными единицами в пределах морфемного инвентаря как одного из параметров порядка в языковой системе. Морфемика и словопроизводство, параметры порядка в языковой системе.

About Equilibrium in the Morphemic Subsystem of Language

Nina F.Klymenko,
O.O.Potebnya Institute of linguistics
at the Academy of Sciences of Ukraine
252001 Ukraine Kiev-1, Hrushevsky str., 4

Phone: (044) 228-26-80

Topical paper

AREA: Quantitative-systemic analysis of units in morphemic subsystem of language

Summary:
Substantiation and verification of the scientific hypothesis about attractive force of root as the factor, which ensures the equilibrium in morphemic subsystem of language. Morphemics, word-generating, parameters of self-organization and self-regulation in language system.

The morphemic subsystem of language demonstrates the action of the laws of word depth, preference and asymmetry. The attractive force of root ensures it's equilibrium, because the root is such a morpheme, which is obligatory for every word and gives the semantic and formal combinability of auxiliary morphemes. The proper attractive force of the root is measured by sum of morphemes' combinations with it, which are possible in certain word position.

The morphemic subsystem of language is determined by interrelations and connections between simple units (morphemes) and complex ones (morphemic structures). Word morphemic structure is the succession of morphemes, which is built on certain rules of combination of separate morphemes' types and classes. We have already written about several laws of word morphemic structure constracting [1]. In this case let us remind those of them, which will help to understand how the equilibrium of language morphemic subsystem becomes apparent on the level of its complex units. It means the analysis of word morphotactic features, which are shown with the help of positional and combinatory indications of morphemes, the possibility of their combination within certain morphemic structures and realization (filling) of every structure by the concrete words of the language.

The analysis of about 160,000 words of Modern Ukrainian language permitted to determine that the nucleus of morphemic subsystem form simple structures, the length of which usually is not higher than 4 morphemes. As a rule, their structure is as follows: PRSS, RF, PRSF (where P - prefix, R - root, S - suffix and F - flexion). Language system gives preference to these structures: they are used with high frequency in the language lexicon. The degree of morphemic structure realization and its usage in the language decrease with the complication of morphemic structure, i.e. with the increase of the number of morphemes.

The limit of word affixal developing is caused by word depth in 7+/-2 units. If morphemic structures have the depth larger than this limit, their degree of realization sharply decreases. Words with 8 or more morphemes don't form even 1 per cent in the Ukrainian language. Theoretically the longest word (the length is measured by morphemes) has such formula: 4P+R+6S or 4P+R+6S+F. It is determined empirically, on the basis of analysis of prefixal and suffixal word development. In the analysed material occur units, which have 4 prefixes and 6 suffixes. Both of them belong to singular constructions, it proves their peripheral position in the language.

Morphemic structure of simple (one-root) word is mostly asymmetric one; it becomes apparent in prevalence of word postpositive part over its prepositive part [2].

On the basis of these data it becomes possible to suppose,that the equilibrium of word morphemic subsystem is regulatedby the law of word depth, which, in its turn, is caused by the limited rate of human operative memory (7+/-2 units). It is shownin such dependence: the degree of morphemic structure realization in the inventory of language units and in the text is thesmaller, the larger is its complexity. Side by side with these indications of word morphemic structure is such indication as the attractive force of the root. Thanks to that the root as the bearer of relative words semantics also plays the main role within every separate word; it holds the morphotactics in those limits, which ensure the strength of language morphemic subsystem. We understand the attractive force of root as the sum of morphemic combinations with it, which are possible in a certain position of the word.

The division of morphemes into the root and auxiliary ones has been accepted in linguistics for a long time. It takes into account the obligatory presence of root in every word as well as another type of its semantics in comparison with the auxiliary, affixal morphemes. The semantics of the latter in some way is given by semantics of the root and becomes apparent only in the context with it. If auxiliary morpheme is polysemantic one, the realization of one or several particular meanings is possible only in the context of their combination with the root.

There is also another condition the fulfilment of which is necessary for recognizing of auxiliary morpheme as independent unit - its obligatory combination with root.

Within the morphemic inventory of language units the specific gravity of roots in ten times (in comparison with suffixes) and even hundred times (in comparison with prefixes and flexions) more than that of affixes. Thus, according to the data of the computer morpheme-word-formative stock of Modern Ukrainian language [3] there are 148 prefixal morphs, 672 suffixal and 18849 root ones.

The analysis of positional and combinatory indications of, for instance, prefixes, proves, that pre-root position is obligatory for all units of this type (primordial and borrowed, old and new). Without it the maintaining of prefix in language morpheme status is impossible. From this point of view the borrowed prefixes are significant, because they show the process of their understanding as separate morphemes in Ukrainian language. Their correlation with words, which have the same or different roots makes possible the determination of such morphemes and their transition into the word-formative formants category of language. For the majority of unproductive borrowed prefixes the pre-root position is the only possible one in Ukrainian language.

It may be possible to suppose, that in the pre-root position the auxiliary (prefixal, in particular) morpheme realizes its main differential meanings, in other positions it expresses these meanings, and more often - only the modified ones.

The measurement of root attractive force with the help of determination of morphemic combinations combinatory force becomes possible due to the account of left-sided and right-sided combinability of each morpheme. The latter may be determined only with the help of frequence-combinatory morphemes' dictionary. For Ukrainian language it is compiled with the help of computer [4]. The dictionary is organized in such way that it permits to take into account all morphemic combinations, which have occured within the words of the computer stock, their positions, combinatory force depending on their word position.

The analysis proves that for all (without any exception) prefixes of Ukrainian language the specific gravity of their combinations with the root in the beginning of the word (position [p]R) is essentially higher than that of their combinations with other prefixes. For the majority of prefixes the position p[p]R, i.e. again pre-root position but already not in the beginning of the word, but after another prefix, is on the second place by the activity of usage. On the third place is position [p]pR, i.e. the beginning of the word before another prefix. The activity of morphemic combinations in other positions is essentially smaller. The prevalence of pre-root prefixal combinations (right-sided valency of prefixes) is obvious. It demonstrates the root attractive force, which "keeps" the morphotactics in equilibrium.

The number of prefixal combinations in pairs, triads proves the insignificant realization of combinatory possibilities of morphemic inventory. With the increase of word prepositive part in each position, with its moving off the root both the number of prefixes, which can occupy the second, third and fourth word positions and the number of various morphemic combinations in each position decreases.

In the same time empiric indices of morphemes' combinability are essentially smaller than of those ones, which are determined theoretically (as the possibility of combination in certain position of all units with each other, for instance, 78 prefixal morphemes with the same 78 prefixes). It is determined that in the second position - [p]pR - only 875 prefixal combinations are realized from 6084 theoretically possible, i.e. 14,4%, in the third position - [p]ppR - from 474552 - 76, i.e. 0,01%, in the fourth position, the last position of prefix before the root for the Ukrainian language we meet 3 morphemic combinations. It proves that combinatory possibilities of morphemes are burdened by semantic combinability with the main word units (roots and word-bases). The semantics of prefixes which are the most distant from the root becomes close to the grammatic one acquiring the meanings of action led to the end.

The economy of usage of morphemic inventory combinatory possibilities is achieved due to the repetition within it not only of separate units, but also of pair ones as well as three- and four-component morphemic combinations. Most of them in answer to the order of word-formation become by analogy the new morphemes. Combinatorial possibilities of some new morphemes, prefixes, in particular, are essentially smaller than that of one-component units. It is evident that such prefixal combinations may be maximum two-component and suffixal combinations of such type may be four-component demonstrating in this case also advantages of post-positive (post-root) part of word.

The frequency of pair combinations of the same functional class units is essentially higher than of three-component and more complex structures. For example, each prefix has several (not more than 3) pairs of prefixal combinations, the frequency of usage of which is essentially higher than of other pairs. These pairs demonstrate the highest semantic combinability between morphemes, i.e. the equilibrium of language morphemic subsystem is ensured by the attractive force of root, and its firmness is connected with economic usage of morphemic inventory, by which the possibilities of separate morphemes combinability between each other are strenthened with multiple application of pairs and triads of these units.

REFERENCES

1. Н.Ф. Клименко, Е.А. Карпіловська. Типи морфемних структур сліву сучасній українській літературній мові // Мовознавство, 1991, N 4, стр. 10-21.

2. Н.Ф. Клименко. Симметрия и асимметрия в морфемных структурах слов современного украинского языка // Морфемология и морфемография. - Владивосток, 1993, стр. 44-55.

3. Н.Ф. Клименко, Е.А. Карпіловська та ін. Морфемно-словотвірний фонд української мови як дослідницька та інформатійно-довідкова система // Мовознавство, 1990, N 6, стр. 41-50.

4. Словник морфем української мови. За ред. Н.Ф. Клименко.- Київ, 1991.

5. Н.Ф. Клименко. Система аффіксального словотворення сучасної української мови.- Київ, 1993, 186 стр.

О равновесии в морфемной подсистеме языка

Клименко Н.Ф.

Резюме:
Обоснование и проверка научной гипотезы о силе притяжения корня ак фактора, выявляющего равновесие в морфемной подсистеме языка.

Морфемика, словообразование, параметры самоорганизации и саморегуляции в языковой системе.

ЛИНГВИСТИЧЕСКОЕ ОБОСНОВАНИЕ ПРОГРАММНОГО СИНТЕЗА СЛОВА (НА МАТЕРИАЛЕ РУССКОГО ЯЗЫКА)

А.А.Кретов, И.Е.Воронина
Воронежский университет
Voronezh State University, Voronezh, Universitetskaya pl. 1
394693 Russia

E-mail: fna@amm.vucnit.voronezh.su

Topical paper

AREA: Computer modelling and automatic generating derivative words in Russian

Сбор и обобщение информации о процессе словообразования требует значительных усилий, что может окончательно заслонить творческую сторону исследовательского процесса. В такой ситуации компьютер может взять на себя рутинную часть работы и сыграть роль тренажера при реализации задач обучения.

Наличие определенных закономерностей порождения русского слова породило две взаимосвязанные задачи: (1) формализации и программного подтверждения выявленных правил и (2) фиксации новых.

Основная цель предпринятого проекта - дать исследователю удобный и разумный инструмент, совершенствующийся по мере накопления нового материала.

Программа реализует следующие возможности.

1) Просмотр и корректировка базового материала, т.е. наборов морфем (корней, приставок, суффиксов и окончаний).

2) Выбор формулы слова, что подразумевает возможную фиксацию отдельных его частей, задание количества приставок и суффиксов. Порождение слова может происходить путем случайного выбора нефиксированных частей или путем полного перебора вариантов.

Самый очевидный случай - полный перебор вариантов, дающий наиболее завершенную картину процесса словообразования.

Синтезированный материал можно просматривать и сохранять в том числе и с необходимыми комментариями).

3) К наиболее интересным и до конца не изученным возможностям относится процесс фильтрации, т.е. прохождение порожденных слов через систему фильтров: фонетический, морфонологический, грамматический, парадигматический, семантический. При этом существует иерархия фильтров.

По желанию пользователя выдается информация о результатах прохождения фильтра, необходимые пояснения в случае отрицательгного результата, или же предусматривается режим выдачи только тех слов, которые прошли фильтр успешно.

Разработка фильтров осуществляется согласно их иерархии: реализация одного фильтра дает материал для формализации другого. Эта часть работы предусматривает совершенствование программы.

Программа предусматривает порождение слов, состоящих не более чем из трех приставок, одного корня и шести суффиксов. В режиме выбора формулы исключена возможность порождения словообразовательной парадигмы (множества слов, образуемых от одного производящего на одном шаге деривации), словообразовательной цепи (множество слов, образуемых за N-ое число шагов деривации, при условии, что на каждом шаге порождается лишь одно слово) и словообразовательного гнезда (множества всех слов, порождаемых от данного производящего и всех его производных на всех шагах деривации).

Если исходить из условия, что шаг деривации с формальной точки зрения состоит в прибавлении к производящему приставки или суффикса, то универсальное (абстрактное) словообразовательное гнездо предстанет в виде сети с четырехугольными ячейками.

Сеть является удобным способом представления материала и одновременно инструментом классификации слов по их морфемной структуре.

Исследование истории развертывния сетей отдельных корней должно дать материал для формулировки закономерностей этого развертывания и в первую очередь - для выявления внутрисистемных запретов, накладываемых на некоторые из путей.

Надо различать принципы создания (развертыания) цепи и принципы наполнения ее узлов речевым материалом.

Развертывание сети осуществляется по правилу: на каждом шаге узел двоится: дополняется приставкой и суффиксом, при этом каждый последующий узел (=слово) содержит на одну морфему больше, чем предшествующий.

На наполнение сети речевым материалом накладываются следующие ограничения:

1) Слово не должно содержать более 7+2 морфем, считая окончание;

2) с каждым шагом возрастает теоретически возможное число слов (комбинаторные возможности образования слов возрастают с каждым шагом);

3) по мере приближения к максимуму (9 морфемам) вероятность наполнения узлов словами, употребляемыми в речи, последовательно убывает, стремясь к нулю.

4) точка перегиба составляющей двух указанных зависимостей находится в интервале от 3 до 6 морфем.

Сеть нужна для особого режима, предполагающего задание фиксированного корня и получение всего материала по всем формулам, предусмотренным сетью. Затем на этом корпусе

отмечается отрицательный материал и выводится положительный рисунок данного корня на канве сети, служащий материалом для последующего этапа анализа, состоящего в сравнении рисунков разных корней и классификации их по сходству-различию.

Сеть является самоподобной рекурсивной структурой. Единицей порождения является цикл, тождественный всей совокупности порождающих возможностей сети на одном шаге деривации. Цикл состоит из Начального (при движении слева направо) члена, конечного члена и M средних членов (M изменяется от 0 до бесконечности, с каждым циклом увеличиваясь на 1; т.е. M:=M+1). Ребро графа символизирует связь производного-потомка с производящим-родителем. У крайних членов цикла такая связь одна (в результате префиксации или суффиксации). У средних - две: в результате префиксации и суффиксации.

Если все узлы сети последовательно пронумеровать сверху вниз и слева направо, то расстояние между производным и его производящим(и) равна i (номеру цикла = шага деривации) - для начальных членов цикла (префиксация), i+1 - для конечных членов цикла (суффиксация) и i+1 и i - для средних членов цикла.

Границы (начальные и конечные члены) цикла вычисляются по формуле Fi=Fi-1 + (Fi-1 - Fi-2) + 1 с F0=1, F1=2 - для начального члена цикла и с F0=1, F1=3 - для конечного члена цикла.

Поскольку каждый номер обозначает узел сети порождения слов, то для каждого номера можно вычислить номер его производящего (производящих) по формуле R (родитель) 1 = P (потомок) - i (номер шага деривации = цикла) - для начального члена (N), R2 = P - (i+1) (K), и R2=P-(i+1); R1=P-i - для средних членов.

Статус члена (начальный, средний, конечный) определяется по формуле Ni <= Pi >= Ki, а границы интервала (начальный и конечный члены), определяются по формуле, приведенной выше.

Из описанной модели порождения слова с необходимостью возникает проблема тождества слова. Применительно к нашему случаю она имеет два аспекта: формальный и генетический. При формальном подходе учитывается только тождество морфемного состава слова - в отвлечении от истории его порождения. При генетическом подходе учитывается последовательность, в которой происходило приращение аффиксов, учитывается история порождения данного морфемного состава. Второй подход содержательно богаче, и его целесообразно удержать.

Тогда тождество слова будет задаваться не только тождеством его морфемного состава, но и его генетической историей, историей его порождения. С учетом истории порождения каждый некрайний узел сети представляет собой совокупность словообразовательных омонимов. И в каждом таком узле окажется столько словообразовательных омонимов, сколько разных путей ведут к нему от вершины графа. Все крайние члены сети будут иметь 0 словообразовательных омонимов, поскольку к ним ведет лишь один путь. Максимально число путей наблюдается у среднего члена цикла, если число узлов в цикле нечетное, или у двух средних узлов цикла, если число узлов в цикле четное. Таким образом, количество словообразовательных омонимов в сети обладает центральной симметрией.

Правила подсчета числа омонимов в узлах сети также рекурсивны. Для средних членов цикла они могут быть выражены формулой: i {i+n [l+n (l+n)...]}, при n:=n+1, где i - порядковый номер цикла; l - число омонимов у члена цикла, находящегося слева; n - переменная, возрастающая на единицу на каждом шаге деривации (с каждым циклом). Формула отражает только левую половину цикла, поскольку вторая половина является ее зеркальным отражением.

С содержательной стороны небезынтересен вывод, что морфемной структурой: три приставки - корень - три суффикса - окончание могут обладать 10 словообразовательных омонимов.

Разумеется, это не значит, что все теоретические возможности реализуются в речи, но сам вывод о наличии такого числа возможностей представляется нетривиальным, поволяя поставить вопрос об исследовании словообразовательной омонимии на принципиально новую основу.

В настоящее время данная программа является исследовательской. Вместе с тем она обладает значительным дидактическим потенциалом, который может быть раскрыт с помощью учебных программ, для которых создается хорошая основа.

Linquistic Substantuation of Programming Synthesis of Word (Using Data of Russian)

Kretov A.A.,Voronina I.E.

Correlations between Semantic, Derivational and Chronological Characteristics of English Adjectives

Leonid A.Kuzmin
Smolensk State Pedagogical Institute
Russia, 214000 Smolensk, Przhevalsky st., 4
Phone: (081-22)-37700

Topical paper.

AREA: Systemic correlations between language features

Summary:
The aim of the present paper is to discuss the results obtained from inter-level search for Pearson correlations of adjectival characteristics in the English language. Basic for the present fragment of adjectival studies are semantic, extra-derivational and chronological characteristics. The discussion is given in a semantic perspective, i.e. the focus is on the correlations of semantic characteristics with the other two categories of properties.

The material under research is confined to a 10% random sampling from the population of A.S.Hornby's Advanced Learner's Dictionary totalling 720 adjectival stems.

The semantic characteristics are represented at the level of energic, informational and ontological classes of meaning. These classes have been pioneered by Prof. Georgiy G. Silnitsky (3,45 - 46) for the Verb. They are now being applied to the Adjective (2,34 -35) both on the basis of empirical data and theoretical postulates on the Verb and Adjective similarition (4;5;6;7).

The energic class (ENERG) embraces the following characteristics: physical (black, hot, wet), physiological (anaemic, hungry, sick), structural and formal (dense, fluffy, square), spacial and dinamic (big, quick, wide).

Inside the informational class (INF) we single out characteristics: emotional and psychic (angry, dreamy, sad), volitional (desirous, eager, reluctant), communication (argumentative, informative, talkative), intellectual (clever, expert, stupid), semiotic (alphabetic, linguistic, readable), sensory (perceptible, sensory, tangible).

The ontological class (ONT) includes the following characteristics: social and ethnic (American, ethnic, rural), evaluative (awful, fine, useful), existential (alive, dead, extinct), quantitative (abundant, numerous, scanty), possessive (deprived, own, tenacious), temporal (annual, daily, late), abstract-qualifying (abstract, special, usual).

The derivational characteristics include generalized parameters, reflacting both the mere fact of "extra-adjectival" word-formation and its means (with certain particularization of the dominant affixal type). In the sum total the present research embraces the following set of "extraderivational" characteristic :any derivative of the given adjective ([DER) (bitter=ness, black=ish, brave adj.- brave v., un=kind);[AFF- affixal derivative (arithmatical=ly, be=dim, im=possible);[PREF - prefixal derivative (ab=normal, dis=able, em=bitter),[SUF- suffixal derivative (abnormal =ity, cautious=ly, Irish=ize); [PREF/SUF - a prefixal as well as suffixal derivative, or so called " omitted-stage", confixal derivative (be =lat=ed, e=long=ate, remorseful=ly, un=remorseful), COMP - a composite stem formed (partially) by the given adjective (bitter- sweet, blue-eyed, fair-haired),[CONV - a stem entering "conversion" with the given adjectival stem (ceremonial a,n; clear a,v; collective a,n); [N - a substantive derivative (absurd=ity, brav=ery, cruel=ty); [V ,- a verbal derivative (en=feeble, en=large, modern=ize); [A - an adjectival derivative (huge=ous, green=ish, ir=relevant);[ADV - an adverbial derivative (angri=ly, bold=ly, wild=ly); a derivative with: AFG - a Germanic affix (dark=le, green=ling, light=ness), AFR - a Romanic affix (comparativ=ist, in=convincible, linear-ity), AFGR - a Greek affix (concret-ize, huge=ous, immortal=ize).

The chronological characteristics are given in terms of the basic periods in the English language history: NE - New English (ascorbic, phonic, republican), ME - Middle English (auburn, ready, safe), OE - Old English (dark, foul, red).

Each of the above-mentioned characteristics is correlated in couples with characteristics of different classes by means of Pearson's correlation analysis procedure. The total number of correlation coefficients computed is 93.

At the level of reliability 95% which is considered sufficient for language studies (1,52 -53) we assume relevant the coefficients obeying the condition Rxy |1,960r|. For the present paper the relevance level is set at |.07|. Of all the coefficients obtained 60 (64,5%) have proved to be relevant.

The relevant correlation coefficients are demonstrated in the table. Statistically irrelevant correlations are represented by a dash.

Table 1.

The correlation of the semantic, derivational and chronological characteristics

	ENERG	INF	ONT	NE	ME	OE
[DER	-.17	.08	.08	-.15	.21	-
[AFF	-.16	.16	.10	-.16	.25	-
[PREF	-.16	-	.08	-.19	.28	-
[SUF	-.14	.19	.13	-.13	.26	-
[PREF/SUF	-.15	-	-	-.20	.28	-
COMP	-	-	-	-.23	.10	.37
CONV	-	-.09	.07	-.18	.11	.17
[N	-.12	-	.09	-.22	.27	-
[V	-	-.07	-	-.17	.13	.14
[A	-.09	-	-	-.16	.22	.16
[ADV	-.21	.16	.07	-.18	.23	-
[AFG	-.19	-	-.40	.25	-	-
[AFR	-.21	-.07	.15	-	.15	-.11
[AFGR	-	-.09	-	-	-	-
NE	.09	-	-			
ME	-.08	-	.08			
OE	-	-.10	-			

The tabulated statistics make it possible to see the peculiarities of system generating connections relative to different characteristics of a certain language level as well as to each of the language levels on the whole.

The energic class is negatively correlated with the general derivational characteristic and such manifistations of extra-adjectival word-formation as affixation , prefixation suffixation, prefixation and

suffixation as a conjoint characteristic ; derivation of the noun, of the adjective of the adverb, with a Germanic affix , with a Romanic affix. A negative correlation is also observed between this semantic class and Middle English period (ME). The only positive correlation of the energic class is that with the New English period.

Thus, from a general systemic viewpoint the energic adjectives are characterized by a relative low word-formation activity and a certain "attachment" of this layer of adjectival semantics to New English stems.

The informational class is positively correlated with extraderivational characteristics , and negatively with CONV,[N, [ADV, AFR, AFGK,as well as with the chronological characteristic OE. In other words this class is positively correlated with the phenomenon of extra-adjectival word-formation per se and with a number of its more concrete manifestations, in particular, with affixation, suffixation, derivation of the Adverb. In the meantime the informational adjectives turned out to be more selective with regard to extraderivation; alongside with the above-mentioned positive correlation they "repel" conversion derivative of the Verb , Romanic and Greek affixes. Worth mentioning is also the negative correlation of the informational class with Old English.

The ontological class is positively correlated with DER, [AFF, [PREF, [SUF, CONV, [N, [ADV, AFR and ME chronological feature. No negative correlations have been obtained for the given semantic class. Thus, the ontological adjectives are highly "disposed" to form derivatives, the conclusion being confirmed by only positive correlations (on contrary to the other two semantic classes) with a member of extraderivational characteristics.

It is also noteworthy that from a chronological viewpoint the ontological class is positively correlated with the Middle English period, thus being a kind of a "correlational antonym" to the energic adjectives in this respect.

Conclusions

1. The energic class is rather clearly opposed to "non-energic" classes from the viewpoint of extra-derivational as well as chronological characteristics; it is characterized by negative correlations with extraderivation and the Middle English period. (See the positive relevant correlations or non-relevant correlations of the informational and ontological meaning with the same characteristics of the other levels).

2. The "non-energic" adjectives, homogenous with respect to the [DER characteristic (the fact of derivation per se) is at the same time characterized by differential properties at the boundary of informational and ontological meanings. (See correlations with CONV, [PREF, [N, [V, AFR, AFGR, ME, OE).

3. The obtained Pearson coefficients are individual for each semantic class whereby correctness of this semantic classification for the Adjective has been confirmed.

4. The most definite correlational boundary is to be drawn between the energic and ontological classes that testifies to their considerable "diagnostic" position in the framework of interlevel connections of the language system elements.

Correlations between extraderivational and chronological characteristics are of subsidiary significance for this otherwise "semanticentric" paper. However, we wish to point out to a quantitative confirmation of lesser word-formation "branching" of New English Adjectives (See negative correlations of the NE characteristic with a majority of extraderivational characteristics) and greater word-formation "involvement" of Middle English and Old English adjectives. At the same time there are good grounds for emphasizing the greatest word-formation relevance of the ME period which is characterized by positive correlations with all the extraderivational features, except AFGR.

The obtained data also make it possible to assume a greater chronological determinedness of word-formation phenomena in comparison with semantic ones. This assumption is based on greater "density" and higher absolute value of relevant correlations observed for chronological characteristics.

References

1. Головин Б.Н. Язык и статистика. М., Просвещение, 1971. - 190 с./ B.Golovin. Language and Statistics. Moscow, "Prosvescheniye Publishers", 1971 - 190 p.
2. Кузьмин Л.А. Формальные и семантические факторы правосторонней синтаксической валентности прилагательного в современном английском языке.//Проблемы синтаксического членения предложения. Смоленск,1992, ñ.34-35 / L.Kuzmin. Formal and semantic factors of right-hand syntactical valency of the adjective in Modern English. In:Problems of syntactic segmentation of the sentence. Smolensk,1992, p.34-35.
3. Сильницкий Г.Г. и др. Соотношение глагольных признаков различных языковых уровней в английском языке. Минск, Навука и тэхніка, 1990 - 182ñ. / G.Silnitsky et al. Correlations of verbal features of different language levels in English. Minsk, Navuka i Technika Publishers, 1990 - 182 p.
4. Anderson J. Adjectives, Datives and Ergativisation. - Foudations of Language, 1969, v.5, p.303-323.
5. Babby L.A. The Deep Structure of Adjectives and Participles in Russian. - Language, 1973, v.49, N 2, p.349-361.
6. Lakoff G. Irregularity in Syntax. - N.Y.: Hold, Rinehart and Winston Inc., 1970. - 207 p.
7. Ross J.R. Adjectives as Noun Phrases. - In: Modern Studies in English: Readings in transformational grammar. Englwood Cliffs (New Jersey): Prenticehall Inc., 1969, p.352-360.

Корреля́ция между семанти́ческими, деривацио́нными и хроноло́гическими характери́стиками англи́йских прилага́тельных.

Кузьмин Л.

Резюме:

Целью настоящей работы является обсуждение результатов, полученных от межуровневого исследования корреляций Пирсона между характеристиками английских прилагательных. Базой для настоящего изучения прилагательных являются семантические, деривационные и хронологические характеристики. Обсуждение дается в семантической перспективе, т.е. концентрируется на корреляции семантических характеристик с другими категориями свойств.

Systematic Characteristics of English Synonyms

Ljalkova I.
Smolensk State Pedagogical Institute
Russia, 214000 Smolensk, Przhevalsky st., 4
Phone: (081-22)-37700

Topical paper

AREA: Quantitative analysis of English synonyms

Summary:

English synonymic verbs are considered in dependence on their morphological structure, derivative complexity, polysyllabic structure, quality of initial vowel and combined use with indirect object, with object clause, with secondary predicate and with adverbial modificator.

According to the criterion of the absence/presence and type of semantic correlations between verbs in present-day English they can be divided into three classes. At the first stage of classification we distinguish between semantically "connected" and "isolated" verbs.

Verbs are semantically connected if they are correlated with other verbs in their lexicographical definitions. Connected verbs, in their turn, are subdivided into a "week" and a "strong" subgroup. The first includes synonymous which are registered in at least one of the dictionaries of synonyms: Crabb's English synonyms. New-York, 1946 or Webster's Dictionary of synonyms, Springfield, Mass., 1973. The verbs of the second subgroup are not registered in any of the dictionaries of synonyms and are defined in the Concise Oxford Dictionary, 4-th ed. Oxford 1956 through other isolated verbs, semantically close to the defined verb, but not exactly synonymous to it.

All the other verbs in the language are isolated. Thus, isolated verbs do not figure in the above mentioned dictionaries of synonyms and are defined in the Concise Oxford Dictionary not by semantically equivalent verbs, but through word combinations.

The present paper has the "polar" classes of synonymous and isolated verbs as its subject in their statistically relevant correlations with characteristics of the following language levels:

1. Diachronical, subdivided into chronological and etymological sublevels;
2. Semantical: semantic types of verbs, verbal monosemy/polysemy;
3. "Introbasal" characteristics, represented within the limits of the verbal base: morphological and phonetic;
4. "Extrabasal" characteristics: derivational and syntactical.

Statistical connections between verbal characteristics are determined by means of the method of correlational analysis (Pearson's criterion). Coefficients not less than |.10| are considered to be relevant. Positive and negative correlations were established on a 5 per cent chance selection of verbs from the Advanced Learner's Dictionary of current English by A.S.Hornby, R.V.Catenby, H.Wakefield, London, 1958 (309 verbs). The choice of this dictionary allows of a comparison of the results obtained by the complex study of the English verb enacted on the basis of the same dictionary by the English department of Smolensk Pedagogical Institute.

1. Diachronical Aspect.

Chronological diachronical features characterize verbs from the origin in the Old English (OE), Middle English (ME) or New English (NE) periods. Etymological characteristics are represented by the Roman (ROM), German (GERM) or Greek (GR) origin of the verbal root and affixes. In the table given below and hereafter irrelevant coefficients are marked by a dash.

Table 1.

Diachronical characteristics of synonymous and isolated verbs

	Synonym.	Isolated
OE	-	-12
ME	.24	-.23
NE	-.25	.23
Rom.root	-	-
Germ.root	-	-
Greek root	-.10	.12
Rom.affix	.20	-.13
Germ.affix	-	-
Greek affix	-.11	-

Conclusions: 1. Synonymous and isolated verbs are characterized respectively by a predominantly Middle English and New English origin.

2. The etymological characteristics of verbal affixes are more relevant than those of verbal roots.

3. The Greek etymology of both roots and affixes is the most diagnostic. A Greek root positively correlates with isolated verbs.

4. Roman affixes have a positive correlation with verbal synonymy.

5. The German etymology of roots and affixes is the least diagnostic.

2. Semantic Aspect.

According to semantic criteria worked out by professor G.Silnitsky [1] verbal meanings are classified into three types: Energetical (ENERG), Informational (INF), Ontological (ONT). Another semantic feature considered in our study is that of verbal monosemy and polysemy.

Table 2.

Semantic characteristics of synonymous and isolated verbs

	Synonym.	Isolated
ENERG	-.21	.20
INF	.19	-.15
ONT	.15	-.15
Monosemy	-.20	.22

Conclusions: 1. Isolated verbs are mainly characterized by the Energetical type of meaning and monosemy;

2. Synonymous verbs are characterized by Informational and Ontological types of meanings and polysemy.

3. Introbasal Characteristics.

The following verbal features are cosidered here: morphological introbasal structure (polymorphemic and derivational structure), phonetic characteristics (initial vowel and monosyllabic structure).

Table 3.

Introbasal characteristics of synonyms and isolated verbs

	Synonym.	Isolated
Polymorph.	.16	-
Prefix	.16	-
Suffix	-	-
Derivation	-.14	-
Polymorph.+ Nonderivation	.30	-.25
Initial Vowel	.11	-.10
Monosyllabism	-.12	.14

Conclusions: 1. Synonymous verbs are characterized by positive correlations with a polymorphemic structure (especially with prefixation), nonderivaton, polysyllabism and an initial vowel.

2. Isolated verbs have no relevant morphological characteristics and are phonetically characterized by monosyllabism and an initial consonant.

Thus, synonymous verbs are characterized by a more complex, isolated verbs - by a simpler formal structure. On the other hand, synonymous verbs tend to a nonderived structure. The combination of these two characteristics (polymorphemic structure and nonderivation) is the main introbasal characteristic of synonymous verbs, as is proved by the highest coefficient of correlation.

4. Extrabasal Characteristics.

The following extrabasal verbal characteristics are represented in Table 4: extraverbal derivation, derivational prefix and suffix, morphological class of the derived forms (verb, noun, adjective) - derivational extrabasal characteristics; transitivity (VT), compatibility with an object clause (CL) and inderect object (Oi), a secondary predicate (2-nd PRED) and an adverbial modifire (MOD) - syntactic characteristics.

Table 4.

Extrabasal characteristics of synonymous and isolated verbs

	Synonym.	Isolated
Extrabasal Derivation	.26	-.26
Deriv.Prefix	-	-
Deriv.Suffix	.29	- .27
Derived Verb	-	-
Derived Noun	.26	-.28
Derived Adjective	.20	-.20
VT	-	-
Oi	.22	-.14
Cl	.14	.13
2-nd Pred	.12	-.13
Mod.	-	-

Conclusions: 1. Synonymous verbs are characterized by a richer and more varied extrabasal valency.

2. In the derivational sphere the most characteristic feature of synonymous verbs is the suffixal derivation of nouns and adjectives.

3. On the syntactic level synonymous verbs are characterized by a heightened compatibility with an inderect object, an object clause and a 2-nd predicate.

4. Isolated verbs are characterized by a reduced extraverbal derivational and syntactic valency.

References

1.Sylnitsky G.G. On the problem of Correlations of Semantical and Formal Characteristics of English Verbs. In: Semantics of the English Verb in Correlation with Characteristics of Different Language Levels. Smolensk, 1988, p.8-9

Систематические характеристики английских синонимических глаголов.

Лялкова И.

Резюме:
Рассматриваются зависимости синонимических глаголов английского языка от их морфемной структуры, деривационной сложности, числа слогов, качества начальной гласной, сочетаемости с непрямым объектом, с дополнительным придаточным, вторичным предикатом и адвербиальным модификатором.

Fractal Presentation of Natural Language Texts and Genetic Code

M.U. Maslov, P.P. Garjaev
Steklov Institute of Mathematics of Russian Academy of Sciences
Russian State Medical University
E-mail: polikarp@logos.msu.su

Topical paper

AREA: Mathematicl modelling in semiotics

Summary:
Generalyzed form of fractal presentation of Natural Language and genetic code texts is suggested. Some applications of it are discussed.

Last decade there are being made intensive attempts to demonstrate the unity of semiotico-linguistic regularities of texts in natural languages and genetic code [Makovsky, 1992; Solovyov, Korolev, Lim, 1992; Jeffrey, 1990; Solovyov, Korolev, Tumanjan, Lim, 1991; Korolev, Solovyov, Tumanjan, 1992; Ratner, 1993.] . However, the formal criteria for this way of studying the languages of genom are not sufficiantely developed. In our investigations [Garjaev, 1993; Garjaev, Vasiliev, Beresin,1991-1992; Garjaev, Chudin, Komissarov, Berezin, Vasiliev, 1991; Garjaev, Gorelik, Moiseenko, Poponin, Chudin, Shtsheglov, 1992; Garjaev, Grigoriev, Vasiliev, Poponin, Shtsheglov, 1992; Agaltsov, Garjaev, Gorelik, Shtsheglov, 1993; Trubnikov, Garjaev, 1993] we have shown that there are different language sustems and subsustems used within the apparatus of heredity of highest biosystems including man. Genom works like a generator of imaginary-simbolic structures of physical and informational levels managing the structure of biosystem. The important part of work of chromosomes is the realisation of DNA commands of the kind of speech setting the strategy of metabolism. It`s not excepted that approaches of the kind would be useful also for analysis of social-genetic regularities, because the human society could be considered as a macroorganism in which the functions of supergenetic `molecule` are fulfiled by the speech and notional formations. In this aspect methods of fractal presentation of DNA sequences gains a special interest in comparison to the same view over the speech sequences in natural languages. Exactly this problem presents the object of the work.

It`s possible to approach the analysis of genetic texts and texts in natural languages from the point of there fractal representation, the so called CGR-presentation of languages. Particularly there was suggested the method of compact and obvious graphic representation of nucleotidic DNA sequences - Chaos Game Representation (CGR) [Jeffrey, 1990]. The procedure of building this representation is described as followes : all the bases are considered to be the points of the square; the first base of sequence is represented by the point of the middle of segment; every next base is represented by the point lying on the middle of the segment wich connects the previous point with the corresponding apex of the square.

The main properties of CGR are shortly formulated as following:

Property 1. Every sequence has the only CGR. Different sequences have different representations.

Property 2. For any point of the square it`s possible to indicate the sequence the last point of representation of which will be on any short distance fixed beforehand.

Property 3. Let`s consider the totality of representations of all possible sequences of the alphabet (A,T,G,C). This totality is a self-similar set with the dimensions of similarity d=2 :

$$d = - \ln N / \ln r(N) = - \ln 4 / \ln (1/2) = 2,$$

where:

N - number of diminished copies of the square which it is covered with,

r(N) - coefficient of scaling smaller than 1, i.e. this set `fills all the square` (like a curve of Peano that also has the fractal dimensions equal to 2).

We could propose the better worked out variant of CGR for the languages with any number of symbols in alphabet, the variant free of defects of representations discribed in items a),b) and satisfying the characteristics 1-3. This approach generalizes the representation used in work [Jeffrey, 1990]. At first, we describe it for the case of the nucleotidic alphabet of symbols.

We divide the square into subsquares (in our case they are 4), then we put the symbols A,T,G,C of the alphabet in correspondence to each of the subsquares. Each of the quarters is similar to the square paper, that`s why it`s possible to reflect it into any of it`s quarters by parallel transfer and scaling with the coefficient r=1/2. This is the one to-one reflection. The centre of the square represents the so-called 'empty' chain of symbols. The reflection of the graphic representation of the previous symbol is a graphic representation of the given symbol; particularly the representation of the first will be represented by reflection of the centre.

It`s obvious that in the case of the 4-symbol alphabet this algorithm causes the same result as the algorithm used in [Jeffrey, 1990].

If we change the coefficient of scaling (for example: r=1/6) and accordingly the number of the diminished copies of the square that cover it (in given case they are 36), we manage to receive the graphic representation of the texts, for example, in Russian putting any of subsquares in correspondence to every of 33 letters of Russian alphabet. Every next symbol, as earlier, determines the transfer into corresponding subsquares; the reflection of the graphic representation of this symbol.

In the works of the biologists [Solovyov, Koroliev, Tumanjan, Lim, 1991; Korolev, Solovyov, Tumanjan, 1992] it was proposed to use CGR in search of the functional parts of DNA. In accordance with every known family of genes there is being built the recognizing matrix, according to the terminology of the authors, 'the mask of the fractal representation of the set' (the mask of FRS). To find the measure of closeness of the given nucleotidic sequence and the family of sequences it is proposed some measure of likeness that uses the mask of FRS. As it`s affirmed in these works the method of masks has the essential merits such as

efficiency (number of operations depends linearly on length of sequence being recognized). The results received by means of this method are also described. We suppose that the method of masks could be found useful in linguistics as well: especially for searching the closeness of texts in natural languages.

For example, the following way of using the CGR in linguistics could be proposed: symbols, but not words, are put in correspondence to the subsquares.

After we put the English auxiliary words "in", "on", "of" and "the" in correspondence to 4 parts of the square, we receive the CGR-reflection of the text `INTEL 387 PROGRAMMER'S REFRENCE MANUAL`.

If the chosen features of the analysed texts (of kind of the English auxiliary words) belong to the essential and informative ones for characterizing the structure of English texts, the received fractal structures could be useful for distinction-identification of various texts in natural languages in different aspects (thematic, stylistic, etc.). Experimental work on this point is in progress.

Фрактальное представление текстов естественного языка и генетического кода

М.Ю.Маслов, П.П.Гаряев

Резюме:
Предложена обобщенная форма фрактального представления текстов естественого языка и генетического кода. Обсуждаются некоторые приложения этого подхода.

The Dynamics of Frequency Structure (Graphic Computer Pattern)

Yu.K. Orlov
Center of scientific information
Presidium of the Academy of Sciences of Georgia
Tbilisi, Georgia
Phone: 67-34-38

Project note

AREA: Computer Modeling of Language

Summary:
Programme allows to analyse textual frequency data and make comparrison them with theoretical formulae and providing different ways of selecting parameters.

The pattern is intended for drawing dependencies of the kind "rank-frequency" in the display of personal computer XT/AT. The programme can:

1) to draw empiric frequency figures "rank-frequency" in given scale and to compare directly them for different texts and selections;

2) to select parameters of theoretical pattern in different variants of selection;

3) to draw theoretical graphics in the same scale as empirical graphics;

4) to carry out theoretical predictions to any other size of a text or a selection;

5) to compare frequency structures of texts of different size with the help of this predictions (what is usually impossible without theoretical pattern.)

Besides, there is a show part of the programme, which illustrates the peculiarities of theoretical formulae and shows different ways of comparison them with the empirical data and different ways of selecting parameters . It is also possible to make a special editorship of pictures for the black-and-white printing.

During selecting parameters it is possible to take into account some peculiarities of frequency structure of texts in Georgian.

Динамика частотной структуры (Графическая компьютерная модель)

Орлов Ю. К.

Резюме:
Программа позволяет анализировать текстовые частотные данные, сравнивать их с теоретическими данными, подбирать параметры

Evolutionary Aspects of a Language as a Natural Classification System

A.A. Polikarpov
Moscow State University,
Department of Theoretical and Computational Linguistics
Russia, 117899, Moscow
Phone: +7 095 939-3178
Fax: +7 095 939-5596
E-mail: polikarp@logos.msu.su

Topical paper

AREA: Systems Theory approach to Language.

Summary:
There are considered problems of viewing Language as a communicative classification system. As a special kind of the whole class, Language inherits all its generic features, but also develops some specific ones. Understanding generic and specific features of Language enables us to model its evolution.

1. There is a great deal of claims in linguistics about wear, degradation of language, (R.Rask, J. Grimm, W. von Gumboldt), loss of grammar and lexicon units and relations during its decay (i.e., in glottochronology) etc. On the other hand, there are many statements about the progress in language, constant rise of its complexity. How do these statements accord with each other? If both processes are found in language reality, how do they interact? May be the view on their kaleidoscopic (J.Vendries), and undirected character of historical dynamics of language is correct? To our mind, only the considerating of cognitive principles of the organization of language makes it possible to access practical role and relative significance of both these processes in language life and volution.

2. The most important element in a cognitive approach to language is its treatment as a natural classification system. The notion of natural classification systems is grounded in the most basic manner in the ideas of general biology (N.A. Bernshtein, P.K. Anokhin, G. Quastler) and general system theory (G.P. Melnikov).

There can be distinguished some places in the Universe where the interaction between the micro-objects (elements) located there is more intensive and constant than the interaction of each of them with something from the outside, which enables this part to be come an integral macro-object or a system opposed to its environment. The global environment of the system, considered only from the point of those currents which are in immediate contact with a system, is a super-system. The system interacts directly with it exists in it, is opposed to it, realizes its exchange potencies.

Global and immediate environment as a whole potentially and actually exceed any system, as their part, in terms of substantual characteristics (mass and energy features of their currents), as well as in terms of structural ones.

3. The interaction between object and environment in a specific cases can be a complementary relationship. An equilibrium-harmonius kind of interaction between them takes place here. More often, however, disparity between currents of an object and its surrounding takes place. It is a reflection of disaccord between the inner organization and exchange potencies of the two counterparts, an object-system and its super-system. The continious absence of accord during some critical period of time leads to the suppression of external activity of the object-system by the activity of a super-system, to the exlusion of an object from the interaction. This leads to gradual decay of some of the object's inner connections and elements and finally - to its destruction. But those objects which are capable of adapting themselves to a supersystem for a critical period of time (i.e. to change their functional, structural and substantial characteristics, so that the interaction with the environment becomes equilibrium-harmonious, supporting them), preserve their place in the global net of interactions.

4. One of the inevitable effects of external and internal adaptation of an object to the environment is the necessity for its self-organization, and even - progressive evolution, i.e. a driff in the direction of complication predetermined by some natural causes. It is necessary to say that the adaptation of an object to its environment is, from some point of view, a reflection of significant properties of its environment. The fuller this reflection is, the larger the margin of safety of the object in further possible relations with the environment. This leads to inevitable evolutionary increase of size, power and/or structural complication of some objects organization in nature. The increase of the extence, mass-energy characteristics of objects leads to their greater stability, their greater independence from changes in the environment. The way of extensive progressive evolution is represented by the evolutionary cosmic succession (elementary particle, atom, molecule, planet, star system, Galaxy...). Another, intensive line of progressive development in nature is represented by the proper biological macromolecular systems and, further, by the proper biological line of the evolution. In this case the adaptation is going on not so much at the expense of accumulation of mass-energy potential (which was mentioned above, provides some objects with greater independence and isolation from the environment), but at the expense of the more various and intensive relations with a supersystem (achieved by accumulation of structural, informational force).

Intensive of evolutionary development is possible only within some specific parts of the Universe, which are screened from the most powerful currents of the environment by some natural factors (for example, on the surface of planets like the Earth, where optimum combination of the inflow of energy from the central star and screening of the hard space radiation by the atmospheric mantle are provided).

5. The results of structural reflection, which are accumulated in special information storages inside the

biological systems are images. Any image, in spite of its directness, includes in itself only a part of those characteristics, whichh are peculiar to the original. I.e., it inevitably reflects structural characteristics not of a single prototype object, but of the whole class of objects or class of object properties, generalizes them. This means, that each image is an abstraction. Constant interaction, comparison of different primary images, detection of their specific and coinciding, constant features leads to the formation of new, progressively more empty abstractions, to the appearance of new classes of more general kinds of objects or more general properties, or aspects of a greater number of objects in nature. This continious classificational activity of living beings leads to the formation of hierarchucaly organized subject and aspectual classificational systems, pyramids of abstractions (G.P. Melnikov) in their reflecting sphere. The classificational system of any individual is constantly changes, because any individual "pyramid of abstractions" is constantly fed with new primary images, new impressions of an individual experience. The dynamics is a consequence of the fact that the situations of vital activity are constantly changing. Individuals also constantly achievely change their situations. Moreover, they constantly exchange life experience. Constant changes of individual classificational systems take place on any level of clasification, - on the primary and on those levels which are built above the primary level. However, the higher the level, the fewer changes are to take place there, because images on higher levels are responsible for more stable characteristics of life situations. Only really revolutionary changes in foundations of life can to a certain extent change the"ideology" of individuals, their view of the most general questions. Even slow, imperceptible, but constant changes in practical life of members of the society, happening in the same direction from generation to generation can eventually lead to noticeable changes on the upper levels of classificational space.

Upper levels of any classification are responsible for the strategy of adaptation of a living being to its environment, lower ones - for tactics.

6. The situations of practical vital activity form practical natural classificational systems. The situations of communacative vital activity form communicative natural classificational systems - Languages. Communicative situations are opposed to practical ones. Practical situations are aimed for getting food, finding cover, running from beasts and so forth; communicative ones are aimed for the "transmission" of information for exchange of experience and opinions, with the partners in communication, (which can be used later, if necessary, in practical acts of vital activity). Communicative situations are interspersed practical ones, serve them depend on them. Classifications elaborated in communicative conditions, first of all, consist of special, communicative abstractions, called meanings. By theactivation of this kind of images, some practical, situationally important image (a sense) can be aroused, to give a hint to the partner on the bars of the resemblance between meaning and sense. Eventually, "tansmission" can be achieved with the help of sending some physical objects-mediators, called signs whose images are also associativelly connected with some meanings in the communicants' minds.

The conjecture of a recipient, i.e. the successful finding of the most suitable sense among those which are similar to meaning used, is possible on the basis of the fact that search of the sense takes place not every where, but in some narrow sense space, which is limited by the current common goals of communicants.

Every sign is connected by preceding communicative practice of the nembers of a given community with some number of meanings. There are fewer signs than meanings. As a result, each sign is polysemantic, at least, potentially. Therefore it possesses some amount of semantic uncertainty. Also, there are significantly fewer meanings than senses. That is why every meaning is to some extent polylsensual, can be characterized by some degree of denotative (referential) vagueness.

The rise of new and the fall ot older signs and meanings, the formation and the constant renovation ot polysemantic structures of signs is the most fundamental basis for general historical dynamics of a natural language system, which is analogous in many respects to the dynamics of any classification system; it depends, as a communicative one, on the processes in some practical classification, but has its specific peculiarities.

7. Regularities of the whole language system dynamics are most vividly illustrated by the tendencies in the historical dynamics of its lexical subsystem.

The lexicon is not only, subsystem which is the most extensive in the number of units and relations, but it also represents those "gates" through which a language is open to the outside. As well, it is an ultimate source of grammatical material (morphemes of different kinds) and grammatical relations. The investigation of the lexicon in this respect is, in the final analysis, a key point for the system analysis of language on the whole.

8. Any really new, i.e. just born, lexical sign is in principle monosemantic. In the course of its further communicative use it can develop a whole set of meanings, can become polysemantic. New meanings appear because sometimes a heuristic connection between some meaning of the word and some sense can consolidate, and the former sense can recieve the status of a new meaning, of a new means of hinting at an other extralinguistic sense.

In the course of the process of new meaning appearance, first, the number of remaining associative valencies of each maternal meaning decreases. Secondly, each new meaning, on the average, will be more abstract than that from which it was born.

This direction of a sign's semantic quality development is predetermined by its survival preference. A word able to acquire more and, more abstract meanings gains an evolutionary advantage over those words which are less able to do this. The advantage consists in a wider referential spheze of a word with more abstract meanings, compared to the case of a word with the same number of meaning but less abstract. It provides a word with a greater stability of existence, greater independence from varions changes in the sphere of senses, higher degree of safety. So, the drift to greater semantic abstractness of a word (and a sign of any other level in natural human language) is one of the global tendencies which one has to be taken into account while modelling language system evolution.

This direction of qualitative development of word meanings further predetermines a decrease of activity of each conclutive meaning in producing new meanings from it, but at the same time it predetermines a increase of the stability, the duration of existence for more recent

meanings. Presumably, the first meanings of words are the most active and unstable. The more recent meanings of words are the most inactive (until disappearence of this ability) and the most stable (but not without limit).

9. On the basis of the given ontological model, and with the introduction of some additional assumptions (for instance, on the possible law of distribution the initial abilities of individual words for some level of activity and stability) the logic of the development of word polysemy over time (by considering the integral result of the processes of production and loss of meanings), as well as the development of its homonymous, synonymous, antonymous, phraseological and other semasiological characteristics, the development of its word-formation categorical potential can be constructed. Besides, it is possible to model the development of the lexicon as a whole, the total lexical population in the course of the time. In this case change of correlation in the number of words of different ages, polysemy, categorical and semasiological characteristics can be explained as depending on the typological status of a language, and also on broadening or narrowing of the sense sphere covered by it.

10. Some other aspects of the problem are discussed in Polikarpov, (1993; 1994), Polikarpov, Kurlov, (1994).

References

1. Polikarpov A.A. A Model of the Word Life Cycle // Contributions to Quantitative Linguistics/ Ed. by R. Koehler, B.B. Rieger. - Dordreht,: Kluwer, 1993.
2. Polikarpov A.A. Zakonomernosti zhiznennogo tsikla slova i evoljutsii jazyka (Regularities of the Word Life Cycle and Evolution of Language).....) // Russkij filologicheskij vestnik. - Vyp. I. (- M.: Moskovsky Litsey), 1994.
3. Polikarpov A.A., Kurlov V.J. Stilistika, semantika, grammatika (Stilistics, Semantics, Grammar) // Voprosy Jazykoznanija, N 1, 1994.

Эволюционные аспекты языка как естественной классификационной системы

Поликарпов А.А.

Резюме:
Рассматриваются эволюционные проблемы, встающие при отнесении языка к классу коммуникативных естественных классификационных систем. Определяются основные особенности любой естественной классификационной системы и выделяются видовые черты коммуникативной разновидности этого класса. Это позволяет моделировать закономерности эволюции систем данного вида.

Menzerath Law for Printed Speech

Vladimir Rykov
Linguistic Institute, Russian Academy of Sciences
Russia, 103009, Moscow
1/12 Semashko str
E-mail: rykov@iling.msk.su

Topical paper

AREA: Laws in Quantitative Linguistics

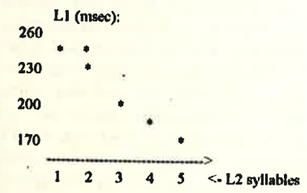
Summary:
Problems of fulfilling of Menzerat Law for units of different levels in written and oral speech are discussed.

Russian linguist Trediakovsky said that written speech is a shadow of the oral speech. Of course, it's not true completely. Modern printed edited speech has many properties independent of oral speech ones. Still, sometimes it is possible to trace footprints of some properties of the oral speech in modern printed texts. It is especially valuable when the investigated phenomenon is controlled by the same brain mechanism. A sample of such a phenomenon is the object of this paper.

Menzerath law was discovered and described in experimental phonetics. The formula of its meaning sounds quite complicated. It means that the length of the component of a speech segment is a function of the length of this segment.

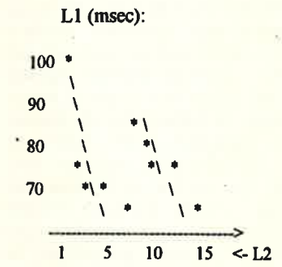
In a simplest form the Menzerath law relation means decreasing of the speech component length with increasing of the length the speech segment which it is part of. In more simple words: the more long is a speech segment - the more it "squeezes" its components.

This form of the Menzerath law is true for describing the decreasing of the sound or syllable length in long words. It can be demonstrated on a simple picture 1 taken from [1]. This "shortening" is because of speech compensation mechanisms which are active in communication process. Certainly it is an illustration of the well known more general Zipf law:



Picture 1. Syllable length L1 (in msec) in 1,2,3,4,5-syllable words (L2)

If the speech segment becomes more and more long the picture does not look so simple. At the beginning the speaker "tries" to follow the simple initial relation. The component length becomes even longer. Then the simple relation "breaks" and repeats itself on a new level. It begins to look like on Picture 2:

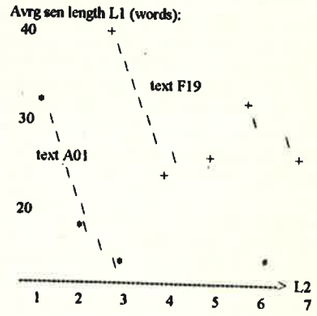


Picture 2. Sound length L1 (in msec) in the words containing L2 sounds.

It means that the speaker breaks too long words mentally in two parts and pronounce them almost separately.

Let us try to trace the phenomena like this in a written speech. The speech segments can be paragraphsm and its component can be a sentence. We will cont the average sentence length L1 (in words) in one-sentence paragraphs, then in two-sentence paragraphs etc.

The results are shown on the Picture 3. They were taken from [2]. The texts were chosen from Brown corpus - so it is easy to check them.



Picture 3. Average sentence length L1 (in words) in paragraphs containing L2 sentences

We can see that the sample text A01 (newspaper) has short - 1,2,3-sentence paragraphs and so it illustrates the simplest form of Menzerath law. The government document F19 has more long paragraphs and shows us the described above phenomenon of mental paragraph breaking into parts.

We have shown one more example that the oral speech phenomena have their shadows in a printed speech. Sometimes these shadows are on a different speech level. But the form of the shadow can be the same - if it is controlled by the same brain mechanisms.

References

1. Gersic S., Altmann G. Laut-Silbe-Wort und das Menzerath'sche Gesetz // Forum Phoneticum 21.- Frankfurt a. M., 1981. - S.115-123.

2. Rykov V. Rhythmic organisation of the prosaic speech. Dissertation. - Moscow: Moscow State University, 1986. - 170 p.

Закон Менцерата для письменной речи

Рыков В.

Резюме:
Обсуждается проблема выполнения закона Менцерата для единиц различных уровней устного и письменного языка.

Stylistic Typology of Texts on the Basis of Quantitative Analysis of Particular Sources of its Content

Svetlana O. Savchuk
Orekhovo-Zujevo Pedogogical Institute
Russia

Topical paper

AREA: Quantitative analysis in stylistics

Summary:
The paper deals with a kind of stylistic analyses based on a number of substantial parameters assumed to be closely connected with basic style-formative factors, which, according to M.M. Bakhtin, are: (1) the speaker's reference to the subject of speach, (2) to the addressee and (3) to another one's statements on the same subject.

With the help of the offered method a corpus of texts belonding to three main styles (scientific, rhetorical and fictional) was analysed.

1. The problem of stylistic typology of texts, the question of choice of grounds for such a typology and for a description of stylistic differences is far from its complete solution; that is why the search of the connection between internal, characteristics of style and its superficial, expression of them becomes very actual. This paper suggests a method of solution of the problem of stylistic differentiation using as the criteria distinguishing styles by a number of parameters, of texts which directly reflects the influence of the internal stylistic formative factors and, therefore, possens an ability to "catch" essential, substantial characteristics of style.

2. As an instrument of revealing stylistic distinctions may serve a concept of "particular sources of content" representing the definite sense positions correlating and coordinating in the process of creation the discourse, form the general content of a text. One can distinguish three kinds of particular sourses of content (PSM).(1) "Focus of attention" is such a subject position, where is placed the most important component of a given substantial fragment of discourse, becoming for a time the center which grouppes around itself the rest elements of contents.(2) "Point of view" is a sense position occupied in the act of communication by its main participants - speaker and listener. There are two aspects of this notion: the authors point of view and the addressee's point of view, being constructed by the author on the basis of his concept of the adressee and his background knowledge. (3) "Speech center" fixes "author's own" or "alien", belonging to another speaker, sense position and permits to find out the speaker's knowledge of cultural context surrounding the subject of speech. The concept of particular sources of content is based on works by M.M. Baktin.

3. The peculiarity of particular sources of content consists in their ability to change in the process of discourse formation. By changing focuses of attention speaker concentrates listener's attention at the necessary elements of contents of the discourse. In the process of view changing, the author's point of view alternates with the point of view of addressee which enables speaker to adapt his position to the position of listener. Changing speech centers speaker includes fragments of another one's statements in his own discourse. This causes the correlation of different senses within one utterance and enriches it with new overtones. The changes of particular sourse of content as communicative actions of speaker, may assume various language forms. At present 25 ways of expressing changes of speech center (SC), 16 ways of expressing changes of points of view (PV) and 13 - of focus of attention (FA) are known.

4. The particular sources of meaning are assumed to be closely connected with basic style formative factors, which, according to M.M.Bakchtin, are "three constitutive moments of discourse": (1) the speaker's reference to the subject of speech (focus of attention), (2) the speaker's reference to the adressee (point of view) and (3) to another one's statements of the same subject (speech center). The process of the alternate changing of PSC can be enterpreted as a manifestation of the essential aspects of style formation. Language forms, realizing these PSC changes are considered to be important characteristics of style.

Thus, the changes of particular sources of content by which the speaker developes his general content, and the language manifestation of this changes represent two sides of the indivisible process of style formation: the internal, substantial aspects of the process (comp.: "form of content") and its external, speech aspect. Hence, it is natural to suppose that the analysis of PSC changes may be used for the purposes of stylistic typology of texts.

5. The verification of diagnostic abilities of PSC was carried out using the material of 45 extracts of 250-400 word usages (the entire co pus contains more than 14 thousand taken from texts belonging to different genres of three main styles (scientific, rhetorical and fictional). It should be mentioned that qualitative variety of texts was also taken into account. So, the analysed corpus includes not only the texts written by "masters of style", which could be "patten" texts, but contains the texts, "unsuccessful" in stylistic respect.

At the first stage of the analysis each text was described by means of PSC changes the borders where these changes occured were marked and the ways means of language lingual realization of these changes hade been studied. At the second step the results of the stylistic analysis were obtained with the help of a number of quantitative and qualitative parameters. These parameters are:

(1) Total number of word occurences in the text; volume of the text - V(t).

(2) Total number of PSC changes in the text (Npsc).

(3) Average number of PSC changes in a unit of the, text volume in 100 words, which is calculated by formula: (Npsc/Vt)*100. Standard

(4) The number of changes of each kind of PSC (focus of attention, point of view, speech center) in a unit of standard text volume (100 words).

(5) Correlation of different kinds of PSM changes; it is calculated in percents of total number of PSC occurences in the text, which is taken as 100%.

(6) Periodicity of PSC changes, which shows how often any kind of PSC occure; it is calculated by formula: Vt/ Npsc

(7) Periodicity of separate changes of each kind of PSC (FA, PV, SC).

(8) Correlation of different language means of expressing PSC changes (their number, kinds, the predominant ways).

The generalized results of quantitative investigation of the entire corpus of texts are presented in the table 1.

6. The results of the investigation have confirmed completely that each of the three main styles - scientific, rhetorical and fictional - is characterized by the peculiar set of various forms of relationship between PSC. On the other hand, the table shows that the texts which belonging to different styles are distinguished by a number of stylistic parameters. Some of them should be noted.

a) The average number of PSC changes as the well as the periodicity of changes (columns 4 and 8) characterizes styles in the following way: the main part of PSC changes has been found out in the texts of rhetorical style, much less number of the changes - in the texts of scientific style. This dependence can be interpreted by the fact that in scientific style multicomponent terminological word combinations functioning as one word are widespread, which fact increases only total vocabulary volume of the text and docs not increase the number of PSC changes which is connected with its real semantic volume.

The correlation of various kinds of PSC changes within one style (columns 5,6,7) demonstrating the difference between scientific, rhetorical and fictional styles, can be explaned by the peculiarity of communicative purposes of each style. Thus, scientific style is charcterized by marked predominance of PV changes, which tells about active coordination of the author of scientific text with his adressee, which is manifested by in using metatextual exertions in the text and varions means of self-commentation. In rhetorical style, was noticed, a considerable predominance of SC changes, which is connected with its controversity: using the connterparts sense position: the author formes his own position and convinces the listener. In fictional style predominance of FA changes has been found out which tells of greater dynamics, highezsense dencity of fictional texts.

7. Besides the general distinctions of three main styles the investigation has revealed some other typological stylistic differences. The comparison of popular science and scientific texts shows significant difference between them in basic parmeters:

a) number and periodicty of PSC changes in popular science style are higher;

b) in correlation of different kinds of PSCs changess the increase of PV changes (column 6) and decrease of SC changes (column 5) are noted in popular science texts in comparison with the average quantity.

However, as it can be noticed, the quantitative characteristics of popular science texts infringing the average statistical indexes, do not draw these texts together with texts of the other styles, i.e., do not erode the borders of the scientific style, but, on the contrary develop the quantitative characteristics peculiar for the scientific style. All this permit to conclude that the so-called popular science style is a variation of the scientific style.

As far as rhetorical style, the analysis of texts belonging to different genres gives a solid foundation for exeption of newspaper texts (informative texts in particular) from the bounds of rhetorical style.

Newspaper texts according tamany important style formative parameters considerably differ from each other, that, firstly, shows the absence of homogeneity of "newspaper style" and, secondly, their divergence from average quantities got for rhetorical style as a whole which tells of cardinal distinction between newspaper texts and rhetorical style.

8. As far as the so-called "pattern" and "unsuccessful" texts are concerned, the comparison of the results of their analysis has shown the following:

a) in all "unsuccessful" texts the decrease of the total number of PSCs changes is strongly in average marked, that accordingly causes the increase of intervals between two next PSCs changes (column 8). It means that in "unsuccessful" texts PSCs on the average changes occur less frequently.

b) yet, it is interesting to note, that the total number and the periodicity of FA changes (columns 7,11) remains at the same level at wich they are in "pattern" texts. This fact gives a reason for considering the average number of FA changes and the periodicity of FA changes as stable characteristics which remain invariable even in stylistically "unsuccessful" texts;

c) generally, (speaking) the changes of quantitative characteristics, marked in "unsuccessful" speaking belonging to rhetorical style, are directed to average quantities of scientific texts; and characteristics of "unsuccessfull" fictional texts are similar to "unsuuccessful" rhetorical texts.

9. The method of analysis described in the paper could serve as a basis for stylistic typology of texts. It's peculiarity consists in using as the criteria of stylistic differentiation not extralinguistic, not formal linguistic, bat exactly stylistic characteristics of text. In spite of specific difference from other directions of stylistic studies the described approach does not contradict them, but may be regasded as an essential supplement to them, confirming some of the results, specifying and clearing up (some ?) other.

There are all reasons to suggest that the offered method of analysis can be effectively used for constructing of more detailed stylistic typology of texts, for revealing of the gente differences between texts, for the analysis of individual style and working out the criteria of its evaluation, fothee fixation of changes in social speech practice and so on.

Table 1

Style realization in text	total number of word occurences in the text	average number of PSC changes in a unit of the text material	Correlation of PSC changes			Periodicity of PSC changes			Correlation "one's own"- "alien" author's		Correlation "author's-addressee's"	
			SC	PV	FA	SC	PV	FA	SC	another's	PV	addressee's PV
Scientific pattern	2107	29,8	18,2%	35,4%	46,4%	3,4	18,5	9,5	7,2	86,2%	13,8%	84,7%	15,3%
unsuccessefull	1239	21,7	11,1%	36,1%	55,8%	4,6	38,7	12,8	8,3	90,6%	9,4%	87,3%	12,7%
popular sc.	1676	31,8	14,1%	38,8%	46,9%	3,1	22,3	8,1	6,7	70,5%	29,3%	86,4%	13,6%
on the whole	5022	28,5	15,5%	36,8%	48,4%	3,5	22,7	9,5	7,3	82,0%	18,0%	85,9%	14,1%
Rhetorical pattern	2329	38,1	36,1%	22,1%	41,8%	2,6	7,3	11,9	6,3	50,5%	49,5%	86,1%	13,9%
unsuccessfull	1144	31,0	22,0%	33,5%	44,5%	3,2	14,7	9,6	7,2	76,0%	24,0%	89,4%	10,6%
on the whole	3473	36,1	32,4%	25,2%	42,5%	2,8	8,6	11,0	6,5	58,9%	41,1%	87,2%	12,8%
Newspaper	2193	25,9	21,8%	21,4%	56,8%	3,9	17,8	18,0	7,0	78,3%	21,7%	93,7%	6,3%
Fictional pattern	2681	33,2	25,0%	15,7%	59,3%	3,0	12,0	19,2	5,1	56,2%	43,8%	86,5%	13,5%
unsuccessfull	927	28,8	15,0%	20,6%	64,4%	3,5	23,2	16,9	5,4	66,0%	34,0%	89,8%	10,2%
on the whole	3608	32,1	23,0%	16,8%	60,2%	3,1	13,6	18,6	5,2	38,7%	41,3%	87,3%	12,7%

Стилистическая типология текста на основе квантитативного анализа частных источников смысла

Савчук С.О.

Резюме

Предлагается подход к стилистическому анализу текстов, основанный на некотором наборе существенных параметров, которые считаются тесно связанными с основными стилеобразующими факторами. Этими факторами, по М.М. Бахтину, являются следующие:

1. отношение говорящего к предмету речи
2. отношение говорящего к адресату
3. отношение говорящего к чужим утверждениям о том же предмете.

С помощью предложенного метода был проанализирован корпус текстов, относящимся к трем основным стилям (научному, риторическому и художественной литературы).

Problems in Measuring and Interpreting Linguistic Variation

Pauli Saukkonen
University of Oulu
Department of Finnish and Saami
PL 111
FIN-90571 Oulu
Finland
Fax +358 81 5533488
E-mail psaukkon@hermes.oulu.fi

Topical paper:

AREA: Methodological problems/ explanation of text phenomena

Summary:
Relative frequencies very crucially depend on the quantity of a unit with which the variables are related. It is important to consider what is a relevant upper category to which a variable belongs. Otherwise interpretation may become difficult or wrong.

Proportional frequencies of linguistic features have usually been counted from the quantity n expressed as the number of words or otherwise from the length of the discourse. But what are the grounds for this very indifferent method? It is problematic to interpret such qiantitative values.

E.g. Douglas Biber (1988) has investigated the variation of 67 linguistic variables in 23 genres related with the length of his texts. He asks: What do these variables really express? Examining the textual dimensions by means of a factor analysis he arrives the following most significant dimensions:

1. Informational versus involved production;
2. Narrative versus non-narrative concerns;
3. Explicit versus situation-dependent reference;
4. Overt expression of persuasion;
5. Abstract versus non-abstract information;
6. On-line informational elaboration.

In many cases there are difficulties in interpretations. His manner of calculation may arise peculiar factors/ dimensions. I would undoubtedly have interpreted his material differently in places.

I have studied Finnish texts in a similar way, but the result is not the best and clearest one.

It would be better to contrast the variables which are comparable and which express in a wide sense the same thing. Such relative frequencies are more relevant which have been counted from the amount of a common, hierarchically upper category. I will give some examples:

1.
- Non-derived and non-compound substantives (stems) as a proportion of their sum, i.e. all substantives; - Derivative and compound substantives;
- Nouns without genitive attributes as a proportion of all nouns, i.e. of all words which can be qualified by a genitive attribute; - Genitive attributes;
- Relative clauses as a proportion of all qualifying structures; - Other qualifying structures;
- Deictic pronouns as a proportion of the total number of nouns and pronouns; - Nouns and other pronouns;
- Anaphoric pronouns as a proportion of all nominal constituents in clauses; - Other nominal constituents;
- Non-coordinated substantives as a proportion of all substantives Copulative co-ordinated substantives.

These variables correlate with each other and they express very clearly a dimension of informational density. The variables on the left side form more analytical and simple constructions while those on the right side are more synthetic and complex counterparts. Biber's dimension 1 can be interpreted as being approximately the same, but it includes problematic variables which could be better located elsewhere. I think that one reason is in calculating.

2.
- Partitive predicatives and partitive objects as a proportio of all predicatives and objects; - Nominative predicatives and accusative objects;
- Indefinite pronouns as a proportion of nouns and pronouns; - Nouns and other pronouns;
- Disjunctive ('or') co-ordinated substantives as a proportion of substantives; - Other substantives;
- Other nouns as a proportion of nouns. - Numerals.

The left hand side variables are more indefinite, inexact, unspecified compared with their more definite, exact and specified counterparts in the right column. These (and some others) form a dimension of specificness. There is no corresponding dimension in Biber's system.

Principially in this way I have managed to make my previous investigations and interpretations easier and more exact. Summarizing my most significant dimensions I can here only mention very briefly the whole system:

(analytic - synthetic (subjective, attitudinal - objective, non-attitudinal (exact, specific - inexact, non-specific (dynamic, proocessive - static; concrete, real - abstract, mental)))).

It differs from Biber's system. In addition to the different calculations there are different materials and different selections of variables, which has naturally a certain effect, but principles of quantification selected for interpretative purposes are anyway in a central position.

Проблемы измерения и интерпретации лингвистических вариаций

Паули Саукконен

Резюме:
Относительные частоты принципиально зависят от количественных характеристик единицы, с которой связаны переменные. Важно учитывать, какая именно категория, которой принадлежит переменная, является релевантной. В противном случае интерпритация может стать сложной или неверной.

On the Problem of "Syntactic Bifurcation" (Polysemy of the English Infinitive)

G. Silnitskaya
Smolensk State Pedagogical Institute
Russia, 214000 Smolensk, Przhevalsky st., 4
Phone: (081-22)-37700

Topical paper

AREA: General problems of syntactic structure

Summary:
The notion of "bifurcation" has been applied for understanding syntactic structure.

The notion of "bifurcation" is of crucial importance in theoretical synergetics. By a "point of bifurcation" is meant an unstable state of system, the further development of which may proceed in several equiprobable directions, the choice between which is determined by fortuitous factors. "In the vicinity of points of bifurcation a decisive role is played by fluctuations, which "choose" the direction to be followed by the system". In the mathematical line we have here "a critical change of parameter resulting in a new solution of equations"[1]. A bifurcation is thus characterized by the presence of "distinctly differentiated states of a system"[2].

The linguistic correlate of the notion of bifurcation is that of polysemy. The presence of several potential meanings of an element of a sentence may be regarded as a manifestation of the instability of the denoted semantic structure, and the contextual elements resolving this polysemy - as casual (fluctuant) factors bringing the system into a new state of equilibrium.

One of the main sources of syntactic polysemy in a sentence is the infinitive. The valential specificity of the infinitive (as of the other nonfinite forms of the verb) is that, while retaining the "right-hand" ("descendent") valency of a finite verb connecting it with the positions of object and adverbial modifier, it loses its "left-hand" ("ascendent") valency connecting it with the subject. On the one hand, this lack of a verbal predicative valency ensures a greater syntactic mobility of the infinitive as compared with a finite verb, allowing it to fulfil the most diverse functions in the sentence. But on the other hand, the grammatical subject in an active sentence usually expressing the logical subject of the predicate, as infinitive, due to the loss of its verbal left-hand valency, is deprived of a grammatically marked subject. There arises a situation of informational indefiniteness where this meaning of subject may in principle be expressed by several different substantives (nouns or pronouns) in the sentence. The resolution of this syntactic indefiniteness, i.e. the identification of the subject of the infinitive depends upon a set of accessory factors varying from one context to another.

A characteristic example of syntactic bifurcation of the type discussed is furnished by an infinite in the specifically English function of attribute. A noun qualified by a transitive infinitive may express either the subject or the object of the verbal action. Cf.:

John has a brother to educate (1)
John has a brother to educate him (2)

In (1) the transitive infinitive is used without a direct object. Its obligatory objective valency is accordingly realized in the syntactic position closest to it - in the qualified noun "brother". The subject of the infinitival action is consequently expressed by the only remaining noun in the sentence, i.e. "John". It should be noted that the infinitive may figure here not only in the active, but likewise in the passive form:

John has a brother to be educated. (3)

In (2) the objective valency of the infinitive is realized by the direct object ("him"), correferential with the noun in the position of subject ("John"), which therefore likewise expresses the object of the infinitive. The only role left for the qualified noun ("brother") is that of subject of the infinitival action.

Analogously in the sentence

John sent Mary a boy to educate (4)

the transitive infinitive, deprived of a direct object, realizes its "natural" objective valency in the position closest to it, i.e. in the qualified noun "boy"; the next proximate noun ("Mary") correspondingly acquires the function of the subject ofthe infinitival action.

On the other hand, in the sentence

John sent Mary his brother to educate her (5)

the infinitive is used with a direct object which realizes its objective valency and is correferential with the noun "Mary"; the noun qualified by the infinitive ("brother") correspondingly expresses here the logical subject of the infinitive.

The valential polysemy of the infinitive may be resolved not only by a direct, but likewise by a prepositional object. Cf.:

John has a friend to speak to (6)
John has a friend to speak to him (7)

The prepositional-objective valency is realized in (7) and non-realized in (6). The qualified noun ("friend") respectively expresses the subject of the infinitive in the first case and its object in the second.

In the absence of an infinitival obligatory adverbial modifier the qualified noun takes upon itself the corresponding semantic function; the function of the ligical subject of the infinitive is accordingly fulfilled by the next proximate substantive, i.e. by the noun in the position of grammatical subject (8) or object (9):

John bought a house to live in (8)
John bought his brother a house to live in (9)

The valential polysemy of an infinitive governed by an adjective is likewise mainly resolved through its objective environment. Cf.:

John is difficult to please (10)
John is eager to please (everybody) (11)

In (11) the transitive infinitive figures without a direct object; the semantic objective function is correspondingly fulfilled by the proximate noun in the position of

grammatical subject ("John"). In (12) the same infinitive is used absolutively; the omitted direct object may in principle be reinstated. In this case the grammatical subject expresses the logical subject of the infinitival action.

Thus, of the synergetic factors resolving the valential bifurcation characteristic of the infinitive the decisive role is played by the presence or absence of its implied right-hand syntactic valency, objective or modificatory. The resolution of the syntactic polysemy of the infinitive may likewise depend upon the meaning of the proximate noun. Thus, in the sentence

John sent Mary some flowers to thank her for her help

(12) the infinitive has a direct object, as in (5); but, due to the incompatibility of the proximate noun "flowers" with the role of logical subject to the infinitival action, this role is fulfilled by the noun "John".

References

1. И.Пригожин. От существующего к возникающему. Время и сложность в физических науках. Москва, "Наука", 1985, с.118,119.
2. Н.Ю.Климонтович. Без формул о синергетике. Минск, "Вышэйшая школа", 1986, с.74.

К проблеме "синтаксической бифуркации" (полисемия английского инфинитива)

Сильницкая Г.

Резюме:
Понятие "бифуркации" используется для понимания синтаксической структуры.

Lexicostatistical analysis of literary characters (problems and approaches)

A.Ya. Shaikevich
Institute of Russian Language
Russian Academy of Sciences

Topical paper

AREA: Quantitative text analysis

Summary:
A technique of variance analysis was tested on Shakespeare's comedies and gave some results for stylistical and social qualification of words and sometimes syntactic constructions (as seen through frequencies of form words). It was used for clustering literary characters.

An objective study of literary characters might prove fruitful in the description of social and stylistic structure of vocabulary, it might be very interesting for discovering hidden structures in literary texts. However, analysis of literary characters is a neglected field in the realm of linguostatistics. This seems natural when both statistical and philogical factors are taken into consideration:

1. The application of statistical tools does not seem promising in view of the very size of individual texts of literary characters (which is small when compared with the size of a literary text as a whole).

2. There are no statistical tools suitable for handling those short texts.

3. Until recently the task seemed unrealistic due to the general dearth of corresponding information.

For some time, however, the situation has been changing radically. Accumalation of literary texts in machine form made possible compilation of huge concordances and dictionaries with rich information on all aspects and fragments of texts (single literary characters included). The appearance of "A Complete and Systematic Concordance to the Works of Shakespeare" (M.Spevack ed., Hildesheim, 1967-1969) is an encouraging sign. In addition to a concordance of all words of the text it gives individual frequency dictionaries to all single characters of the plays.

It is hoped to obtain comparable statistical information as a by-product of a project of 'Dostoevski's Dictionary' now carried on at the Institute on Russian language in Moscow. (The illustrative examples in the present paper are drawn from both Shakespeare's comedies and Dostoevski's prose works).

Of course, it is too early to speak of the results. Our aim is to point out some problems and perhaps to show some ways of their solution.

One way of circumventing difficulties associated with the small size of individual texts is to operate with groups of characters, classified on the basis of some meaningful principles (such as gender, age, social position), and with groups of texts classified according to genre, form of expression (verse or prose) or time of creation. The apparatus of variance analysis appears quite adequate in this case.

This technique was tested on Shakespeare's comedies and gave some results for stylistical and social qualification of words and sometimes syntactic constructions (as seen through frequencies of form words). When those results were taken into account it was possible to correct the raw frequency data and to analyze quantitavely the 117x117 character-to-character matrices. The corresponding graphs of links between characters showed some interesting clusters sometimes interpreted as social groups (e.g. monarchs, servants) and sometimes as functional groups (e.g. young lovers, middle-aged 'serious' characters).

The opposite approach is being tested on the characters of Dostoevski. The starting point of analysis is compilation of individual lists of lexical markers (words, whose frequencies differ significantly from the mathematical expectation calculated on the basis of word frequencies in the text as a whole). As a rule such a list contains a few dozen words (or word combinations). Of these some words serve exclusively as a means of individualization, some serve as markers of a social group and some as stylistic markers. Perhaps half of them are related to this or that episode of the narration and may form a good base for constructing the 'plot space' of characters. For some of the characters it is possible to organize lexical characters into an individual semantic space.

Statistical and philological problems of further intertext comparison of literary characters are discussed.

Лексико-статистический анализ литературных образов (проблемы и подходы)

Шайкевич А.Я.

Резюме:
Используется статистическая техника группировки и разграничения действующих лиц в литературном произведении.

Inter-Level Correlations of Etymological Types of English Verbs

M. Soldatenkova,
Smolensk State Pedagogical Institute
Russia, 214000 Smolensk, Przhevalsky st., 4
Phone: (081-22)-37700

S. Sergutina
Smolensk State Pedagogical Institute
Russia, 214000 Smolensk, Przhevalsky st., 4
Phone: (081-22)-37700

Topical paper

AREA: System analysis in linguistics

Summary:
The inter-level dependencies between etymological, chronological, morphemic, morphological, phonetic, semantic and syntactic features of Enlish verbs are analysed.

This paper is correlated to a discussion of statistically valid relations between verbal features of different language levels in English.

The basic characteristic under analysis is the etymology of the root:

- Germanic root (GERM): smoke, spring, frighten
- Romanic root (ROM): doctor, aviate, refract

In the research project reviewed in this paper Pearson's statistical criterion was applied to a database comprising 5% selection of English verbs (309 lexical entries) recorded in A.S.Hornby's "Oxford Advanced Learners' Dictionary" 1980. Coefficients not lower than |.08| are regarded as statistically relevant.

The percentage of the verbs with a Germanic root is 39%, with a Romanic root - 53% of the selected list. These etymological features are correlated with diachronic, semantic and formal verbal characteristics.

Formal features are subdivided into the introbasal and xtrabasal subtypes. Inbasal formal features are represented within a single verbal base. These comprise morphological and phonetic features. Extrabasal formal characteristics (extraverbal derivation and syntax) transcend the limits of a separate verbal base.

I. The Correlation of Diachronic and Etymological Features of English Verbs

The diachronic features represent the first appearence of a verb in English in terms of the Old English (OE), Middle English (ME) and New English (NE) periods.

Table 1.
The correlation of the diachronic and etymological features

	OE	ME	NE
GERM	.28	–	-.14
ROM	-.30	.15	–

(Statistically irrelevant correlations are represented by a dash)

Conclusions: 1. Verbs of Old English origin are positively correlated with Germanic roots.

2. Verbs of Middle English origin have a positive correlation with Romanic roots.

II. The Correlation of Semantic and Etymological Features

On semantic grounds verbal features are grouped (a) on the basis of the generalized notions of "energic" (ENERG), "informational" (INF) and "ontological"(ONT) verbal meanings [1] and (b) on the criterion of monosemy (MONOSEMY)/polysemy.

Table 2
The correlations of semantic and etymological features.

	ENERG	INF	ONT	MONOSEMY
GERM	.26	-.15	-.22	–
ROM	-.22	.17	.21	.08

Conclusions: 1. Energic meanings are positively correlated with Germanic roots.

2. Informational and ontological meanings are characteristic of Romanic roots, which have a positive correlation with monosemy.

III. The Correlation of Introbasal and Etymological Characteristics

The morphological introbasal characteristics represented in this paper reflect the morphemic and derivational structure of verbal basis: "monomorphic" (1-morph), suffix (SUF), prefix (PREF), "internal derivation" (DER]), "internal conversion" (CONV]).

The phonetic characteristics comprise: "initial vowel" (VL), "monosyllabic" (1-SYL).

Table 3.
The correlation of introbasal and etymological features

	1-morph	PREF	SUF	DER]	CONV]	1-SYL	VL
GERM	-.30	-.25-	.13	.10	.14	.19	-.25
ROM	.25	.29	.14	–	-.11	-.18	.22

Conclusions; 1. Verbs with Germanic roots are positively correlated with formal simplicity: they are characterized by a monomorphemic and monosyllabic structure.

2. Verbs with Romanic roots are characterized by the presence of prefixes and suffixes in their structure. On the phonetic level Romanic verbs are negetively correlated with monosyllabism and positively - with an initial vowel.

IV. The Correlation of Extrabasal and Etymological Features

1) The derivational extrabasal verbal features comprise: "extraverbal derivation" ([DER), "derived noun" ([N), "derived adjectives" ([A), "prefixal extraverbal derivation" ([PREF), "suffixal extraverbal derivation" ([SUF).

2) The syntactic extrabasal verbal features: transitivity (Vt), combinability of the verb with the syntactic positions of inderect object (Oi), object clause (CL) and adverbial modifier (Mod).

Table 4.
The correlation of extrabasal and etymological features

	[DER	[N	[A	[PREF	[SUF	Vt	Mod	Oi	CL
GERM	-.10	-.11	-.12	–	-.08	33.17	.24	-.14	-.14
ROM	.23	.23	.17	–	.22	33 --	-.26	.14	.13

Conclusions: 1. Verbs with Germanic roots are negatively correlated with extraverbal derivation and positively with transitivity and combinability with an inderect object.

2. On the other hand, verbs of Romanic origin are positively correlated with extraverbal derivation especially with suffixal formation of adjectives and nouns. On the syntactic level these verbs are characterized by positive correlations with the inderect object and object clause.

References

1. Сильницкий Г.Г. и др. Соотношения глагольных признаков различных языковых уровней в английском языке. Минск. Навука и тэхніка. 1990.

Межуровневые корреляции этимологических типов английских глаголов

Солдатенкова М., Сергутина С.

Резюме:
Анализируются взаимозависимости между характеристиками английского глагола различных уровней - этимологическими, хронологическими, морфемными, морфологическими, семантическими и синтаксическими.

The Stairway of Subsustems

Ludek Hrebicek
Pod vodarenskou vezi 4
182 08 Prag 8
Czech Republic

Topical paper

AREA: Systems Theory Approach to Language.

Summary
It is put a quastion, how all thinkable language levels are mutually coordinated in a text. It is possible on the basis of generalization of the case of interaction between two neighbouring levels (Menzerath Law) to the case with unlimited number of levels.

One of the most complicated tasks in theoretical linguistics is the explanation of the way in which different language subsystems cooperate. Generally, it holds that lower subsystems consist of units functioning as constituents of higher constructs. This is a sort of coherence forming the stairway of language levels (or subsystems). 'Construct' and 'constituent' are general concepts which should be applicable to the whole string of units in which constructs represent also constituents of some higher constructs. In the present contribution we put the question, how all thinkable language levels are mutually coordinated in a text.

For arbitrary two levels constituting the relation of a language construct (representing the higher level) and its constituents (forming the lower level), this problem is solved by the Menzerath-Altmann (MA) law; see, for example Altmann (1980) or Altmann & Schwibbe (1989).

The model resembling a stairway of levels is, in fact, a synergetic system containing a great number of cooperative subsystems; it was elaborated in natural sciences and introduced into linguistics by Köhler (1986). This author also published contributions to the analyses of the MA law, see, for example, the above quoted work by Altmann & Schwibbe (1989, 108-112).

In the present paper we try to enlarge the conception of two levels, forming constructs and their constituents, to certain expression of the MA law which is valid for an arbitrary number of levels greater than two. It is evident that the amalgamation of subsystems represents a special sort of text cohesion, - a cohesion free of a direct semantic interpretation (always required in the analyses of text references) and representing a phenomenon implanted deeply in each language edifice. It can scarcely be supposed to be the property of a certain language; it seems to be a phenomenon reaching up to human mind and its biological carrier.

The formula of the stairway

The MA law says that *the longer a language construct in the number of constituents the shorter the mean length of constituents measured in some lower units*. When x is a construct and y is a constituent, their relation is

$$y = A x^{-b} \quad , \tag{1}$$

where A and b are parameters.

When we try to proceed to more than two levels, which are involved in (1), we can use indices; constructs and constituents are then denoted as follows:

x_1 constructs on the highest level assumed (for example, sentence length in the number of words);
y_1 constituents of the preceding level (for example, word length in the number of syllables);
x_2 constructs on the lower level (for example, word length in the number of syllables);
y_2 constituents of the preceding level (for example, syllable length in the number of phonemes).

Obviously, $y_1 = x_2$. The same identity can be used for an arbitrarily long string of levels:

$$\begin{aligned} x &\rightarrow y \equiv x_1 \\ x_1 &\rightarrow y_1 \equiv x_2 \\ x_2 &\rightarrow y_2 \equiv x_3 \\ x_3 &\rightarrow y_3 \equiv x_4 \\ &\dots \end{aligned}$$

Formula (1) can be rewritten in the following form:

$$x_2 = A x_1^{-b} \quad ,$$

or

$$x_1 = \left(\frac{A_1}{x_2}\right)^{\frac{1}{b_1}} \quad , \tag{2}$$

respectively. The similar formula can also be written for x_2, x_3, etc., and all these formulae can be condensed as follows:

$$x_1 = \left(\frac{A_1}{x_2}\right)^{\frac{1}{b_1}} = \left(\frac{A_1}{\left(\frac{A_2}{x_3}\right)^{\frac{1}{b_2}}}\right)^{\frac{1}{b_1}} = \ldots \qquad (3)$$

The same structure can be expressed in the logarithmic form:

$$\log x_1 = \frac{1}{b_1} \log A_1 - \frac{1}{b_1\, b_2} \log A_2 + \frac{1}{b_1\, b_2\, b_3} \log A_3 - \qquad (4)$$

$$- \frac{1}{b_1\, b_2\, b_3} \log x_4 = \ldots$$

In this way the number of levels can be infinitely enlarged.

The structure of the MA law in the form of formula (1) obviously corresponds to the facts observed in natural languages, this already was many times proved. The lowest level which we are able to incorporate, with respect to the contemporaneous knowledge, into our assumptions, is the level containing components of sound spectra. Nevertheless, the signals running articulation are carried by neurons; this biological system is doubtlessly structured into a number of levels which probably form a string of some constructs and constituents. So far, this part of the supposed string is not described in linguistics and the present author's notes concerning this topic cannot be more than a pure speculation.

The highest levels seem to be also numerous, or presumably infinite; let us note the possibility to interpret a certain semantic unit (expressed, for example, by a word) through a text; and this new text is also composed from semantic units which can be semantically interpreted, and so forth. This confirms the complicated structure of a semantic system which is carried, among other carriers, also by natural language.

Syntactic levels can be taken as representatives of language levels taken from a middle part of the entire assumed spectrum of levels. Sentence structure can be displayed into a hierarchy which cannot be conjecturally limited in advance. On each of these syntactic levels we also are dealing with the relations of constructs and constituents, this fact was already proved.

The existence of such intermediate levels is testified by linguistic

intuition with its terminology of 'morphophonology', 'morphosyntax', 'lexicosyntax' and conceivably other branches combining different intuitively (i.e. without the testing by the MA law) stated language levels.

The MA law was used for constituting one supra-sentence level of the units provisionally called 'sentence aggregates', see Hřebíček (1989, 1992, 1993) and Schwarz (1992). These aggregates of a text represent language constructs having sentences as their constituents; each sentence of an aggregate contains a given lexical unit. The existence of aggregates was proved on Old Ottoman, modern Turkish and German texts, and on one English text. Thus it can be asserted that *the longer an aggregate construct in the number of sentences the shorter its mean sentence length.* The structure of aggregates appears to be certain representation of a complete semantic structure of each text.

Application

Formulae (3) and (4) were proved on Turkish texts; here the results obtained from one of these texts are presented.[1] The observed data concerning variables x (= the number of syllables in a word, i.e. constructs) and y (= the mean number of phonemes in syllables, i.e. constituents) are presented in Table 1 together with the values of parameters A and b as well as with the expected length of mean constituents Y. The other Tables are arranged in a similar way. Their content is as follows:

Table 2: x = the number of morphemes in a word,
y = the mean number of phonemes in morphemes.

Table 3: x = the number of words in a sentence,
y = the mean number of syllables in words.

Table 4: x = the number of words in a sentence,
y = the mean number of morphemes in words.

Table 5: x = the number of sentences in an aggregate,
y = the mean number of words in sentences.

Syllables and morphemes are treated as parallel constituents of words. This was proved in an earlier experiment by testing the mutual relation of these two constructs; the conclusion was that morphemes are constituents of

[1] From the political commentary written by Yaşar Nabi ('1967 ye Toplu bir Bakış.' In: 'Varlık Yıllığı 1968.' Istanbul, Varlık, 1967) the part 'Ölünmüs Dünyamız' was chosen.

syllables and, at the same time, syllables are constituents of morphemes in the sense defined by the MA law. Consequently, the stairway proved has two wings which are connected at their ends: syllables, as well as morphemes are constituents of words.

Let us rename the above mentioned levels and present the mean length of the respective constructs by the following symbols (their length being given in the number of units of the lower level):

x_1 = sentences,
x_2 = words,
x_3 = syllables/morphemes. (The length of syllables/morphemes is given in the number of phonemes.)

The observed values of the respective variables and their parameters are as follows:

x_1 = 12.21 (the mean number of words in a sentence);
A_1 = 3.33, b_1 = -0.0423 (the parameters of the relation between words and syllables);
A_2 = 2.47, b_2 = -0.0420 (the parameters of the relation between syllables and phonemes);
x_3 = 2.34 (the mean number of phonemes in a syllable).

Parameters A and b, in the present test as well as in the following ones, were estimated from the observed data of the respective Tables. The negative values of each parameter b must be substituted with positive signs. However, when the above data are substituted for the respective variables of formula (4), we do not obtain too satisfactory result. For this reason we must better estimate the value of x_3 and replace it by the value 2.358821882. Then we obtain:

$$\log x_1 = \frac{1}{b_1} \log A_1 - \frac{1}{b_1 b_2} \log A_2 + \frac{1}{b_1 b_2} \log x_3 \doteq$$

$$\doteq 12.35 - 221.04 + 209.78 = 1.09 \doteq \log 12.21$$

The above substitution of the value of x_3 indicates, how sensitive the variables involved are also on higher decimal places. Nonetheless, the mutual correspondence of three levels is evident.

The analogical string including morphemes instead of syllables operates

with the following values:

x_1 = 12,21 A_1 = 2.83 A_2 = 4.06
x_3 = 2.80 b_1 = -0.0264 b_2 = -0.3811

Then we obtain:

$\log 12.21 \doteq 17.11 - 60.48 + 44.46 \doteq 1.09$

When one higher level (namely, the level of sentence aggregates) is added to this string, we substitute the following variables and their values for those in (4):

x_1 = 1.39 (the mean length of aggregates in the number of sentences);
x_4 = 2.36 (the mean length of syllables in the number of phonemes);
A_1 = 15.47; A_2 = 3.33; A_3 = 2.47;
b_1 = -0.0083; b_2 = -0.0423; b_3 = -0.0420.

The substitution results in:

$\log 1.39 \doteq 143.31 - 1488.06 + 26631.14 - 25286.24 \doteq 0.15$

In this computation the value of x_4 was changed to the value 2.359746399. Such changes can be treated as corrections of earlier estimates obtained by observations which always contain errors. In this way also the other variables and parameters can be corrected on the basis of the MA law.

The parallel variant with morphemes instead of syllables is as follows:

x_1 = 1.39 (the mean length of aggregates in the number of sentences);
x_4 = 2.80 (the mean length of morphemes in the number of phonemes);
A_1 = 16.00; A_2 = 2.83; A_3 = 4.06;
b_1 = -0.0083; b_2 = -0.0264; b_3 = -0.3811.

The resulting equation is:

$0.143 \doteq 143.31 - 2061.82 + 7287.16 - 5368.50 \doteq 0.15$

We can conclude that the structure expressed by the MA law in (1) is a part of a logarithmic multinomial containing more than three members, as it

Table 1

Words (x) and syllables (y)

x	z	s=xz	p	y=p/(xz)	$Y=Ax^{-b}$
1	43	43	106	2.47	2.47
2	120	240	570	2.40	2.40
3	154	462	1083	2.34	2.36
4	97	388	899	2.32	2.33
5	46	230	536	2.33	2.31
6	12	72	164	2.28	2.29
7	3	21	43	2.05	
8	1	8	18	2.25	
Σ	476	1464	3424	-	-

A = 2.4669; b= -0.0420 (both estimated from the data for x = {1 to 6}, for which z is sufficiently high; the same approach is applied also in the following Tables).

x = word length in the number of syllables;

z = the number of words;

p = the total number of phonemes.

Table 2

Words (x) and morphemes (y)

x	z	m=xz	p	y=p/(xz)	$Y=Ax^{-b}$
1	107	107	431	4.03	4.06
2	151	302	947	3.14	3.12
3	90	270	742	2.75	2.67
4	71	284	662	2.33	2.39
5	43	215	461	2.14	2.20
6	12	72	151	2.10	2.05
7	2	14	30	2.14	
Σ	476	1264	3424	-	-

A = 4.0589; b = -0.3811.

x = word length in the number of morphemes;

z = the number of words;

p = the total number of phonemes.

Table 3

Sentences (x) and words measured in syllables (y)

x	z	w=xz	s	y=s/(xz)	$Y=Ax^{-b}$
3	2	6	22	3.67	3.18
4	2	8	30	3.75	3.14
5	3	15	46	3.07	3.11
6	1	6	16	2.67	3.09
7	4	28	82	2.93	3.07
8	2	16	42	2.63	3.05
9	1	9	28	3.11	3.04
10	1	10	27	2.70	3.02
11	3	33	107	3.24	3.01
12	3	36	110	3.06	3.00
13	2	26	79	2.19	2.99
14	3	42	133	3.17	2.98
15	1	15	44	2.93	2.97
16	1	16	48	3.00	2.96
17	2	34	97	2.85	2.96
18	2	36	110	3.06	2.95
20	1	20	60	3.00	2.94
22	2	44	137	3.11	2.92
24	1	24	76	3.17	2.91
26	2	52	170	3.27	2.90
Σ	39	476	1464	-	-

A = 3.3330; b = -0.0423.

x = sentence length in the number of words;

z = the number of sentences;

s = the total number of syllables.

The concordance of the total distributions of y and Y was tested with the help of the non-parametric Wilcoxon T. The testing criterion T = 87.5 > $T_{0.05}$ (20) = 52. The hypothesis concerning the incongruence of the two tested distributions must be rejected.

Table 4

Sentences (x) and words measured in morphemes (y)

x	z	w=xz	m	y=m/(xz)	$Y=Ax^{-b}$
3	2	6	19	3.17	2.74
4	2	8	26	3.25	2.72
5	3	15	36	2.40	2.71
6	1	6	14	2.33	2.69
7	4	28	70	2.50	2.68
8	2	16	36	2.25	2.67
9	1	9	26	2.89	2.67
10	1	10	25	2.50	2.66
11	3	33	91	2.76	2.65
12	3	36	98	2.72	2.65
13	2	26	72	2.77	2.64
14	3	42	117	2.79	2.64
15	1	15	41	2.73	2.63
16	1	16	37	2.31	2.63

17	2	34	78	2.29	2.62
18	2	36	91	2.53	2.62
20	1	20	54	2.70	2.61
22	2	44	121	2.75	2.60
24	1	24	70	2.92	2.60
26	2	52	142	2.73	2.59
Σ	39	476	1264	-	-

A = 2.8254; b = -0.0264.

x = sentence length in the number of words;

z = the number of sentences;

m = the total number of morphemes.

Wilcoxon T = 102 > $T_{0.05}$ = 52. The hypothesis concerning the incongruence of *y* and *Y* must be rejected.

Table 5

Aggregates (*x*) and sentences (*y*)

x	z	i=xz	w	y=w/(xz)	$Y=Ax^{-b}$
1	190	190	2980	15.68	15.72
2	39	78	1210	15.51	15.68
3	15	45	618	13.73	15.66
4	11	44	760	17.27	15.56
5	7	35	517	14.77	15.63
6	1	6	128	21.33	
7	2	14	232	16.57	
8	1	8	101	12.63	
9	1	9	142	15.78	
10	1	10	152	15.20	
11	1	11	174	15.59	
12	1	12	164	13.67	
14	1	14	235	16.79	
Σ	271	376	-	-	-

A = 15.4709; b = -0.0083 (both estimated from the values for *x* = {1 to 5}).

x = the length of aggregates in the number of sentences;

z = the number of aggregates;

w = the total number of words.

corresponds to (1).

The question arises whether for arbitrary two neighbouring multinomials having the number of members $m - 1$ and m certain transformation can be defined, from which a higher level and its properties becomes evident. For example, let us take the level called 'text' as a language construct with sentence aggregates as its constituents; then also text should be considered as a constituent of some higher level, probably of something what is designated as semantics or semantic system. The answer to this question, however, is the target of hereafter experiments.

References

Altmann, G. (1980): Prolegomena to Menzerath's law. *Glottometrika 2*, 1-10.

Altmann, G. & Schwibbe, M.H. (1989): *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen.* Hildesheim-Zürich-New York, Olms.

Hřebíček, L. (1989): The Menzerath-Altmann law on the semantic level. *Glottometrika 11*, 47-56.

Hřebíček, L. (1992): *Text in communication: supra-sentence structures.* Bochum, Brockmeyer.

Hřebíček, L. (1993): Text as a construct of aggregations. In: R. Köhler & B.B. Rieger (eds.): *Contributions to quantitative linguistics.* Dordrecht-Boston-London, Kluwer, 33-39.

Köhler, R. (1986): *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum, Brockmeyer.

Schwarz, C. (1992): Zur Verteilung von Aggregaten in Texten. Ruhr-Universität Bochum. [Unpublished seminar work.]

Лестница подсистем

Гржебичек, Людек

Резюме:
Ставится вопрос о том, как все мыслимое многообразие языковых уровней взаимно координировано в тексте. Это осуществляется за счет обобщения случая взаимодействия двух уровней (закон Менцерата) до случая с неограниченным числом уровней.

Опыт квантитативно-системного исследования словаря гиперлексем русского языка

Каримова Г.О.
Чимкентский государственный педагогический институт
Чимкент, Казахстан
E-mail: polikarp@logos.msu.su

Доклад

ТЕМАТИЧЕСКАЯ ОБЛАСТЬ: Квантитативно-системная лексикология

Резюме:
Рассматриваются принципы выделения гиперлексемы как единицы лексческой системы языка. Излагаются принципы создания словаря гиперлексем русского языка. Оцениваются его некоторые квантитативно-системные характеристики

Современной лексикографии свойственна тенденция к постоянному расширению круга словарей различных типов и назначений (ср.: словари ударений, усилительных словосочетаний, омонимов, паронимов, слитно или раздельно, прописная или строчная и т.д.). Это объясняется, в первую очередь, стремлением современной науки о языке представить в форме словаря различные лингвистические описания и результаты (1).

Словарь гиперлексем представляет собой еще одну попытку "ословаривания" результатов лингвистического анализа, а именно результатов исследования одной из важнейших проблем лексикологии - проблемы выявления и изучения системной организации лексики.

Основной единицей словаря является гиперлексема, которая образуется путем объединения ряда словообразовательно связанных лексем в лексические парадигмы. Группировка лексем в гиперлексемы осуществляется на принципе регулярности. Это значит, что каждая из лексем, входящих в гиперлексему, имеет хотя бы с одним из членов данной гиперлексемы такие и только такие формально-семантические отношения, которые являются регулярными, т.е. повторяются по крайней мере в еще одной паре словообразовательно связанных лексем.

Вода, водица, водичка, водный, водяной, водянистый, водянисто, водянистость, водянеть, безводный, безводье, обезводеть, надводный, подводный. Перед нами ряд слов, обладающих общей корневой морфемой -вод- которая, выполняя функцию семантического ядра, вместе с тем сама по себе не имеет ни структурной, ни семантической законченности. Помимо общего корня -вод- , каждое из перечисленных производных слов содержит в своем составе различные аффиксы (приставки, суффиксы, окончания), необходимые для семантической завершенности и грамматической оформленности такой единицы языка, как слово. Присоединение того или иного аффикса к той или иной производящей основе влечет за собой различную степень переосмысления лексического значения исходного слова в результате взаимодействия значений основы и аффиксов.

Суффикс может просто переводить производящее слово из сферы одной части речи в другую, сохраняя при этом в производном слове хотя бы часть лексических значений производящего: вода - водный, вода - водяной, водяной - водянеть, безводный - безводье. Приставка и суффикс способны в достаточной степени регулярно модифицировать значение исходного слова (добавляя к основному значению некоторый дополнительный признак - единичности, уменьшительности, собирательности женскости, подобия, отрицания и т.д.), не уступая в этом смысле регулярности в словоизменительных парадигмах: вода - водица, водичка; водный ("содержащий воду") - безводный ("лишенный воды") (2). В обоих случаях наблюдается регулярное преобразование лексических значений, по которым соотносятся производящее и производное слова (ср.: вода:водица = лужа:лужица; вода:водичка = сестра:сестричка; водный:безводный = земельный:безземельный).

В словообразовательной паре вода - обезводеть словообразовательные форманты обез- и -е- не просто видоизменяют лексическое и/или общекатегориальное значение производного слова, но способствуют появлению совершенно иного понятия по сравнению с тем, которое выражается производящим словом. В данном случае основанием для включения лексемы обезводеть в рассматриваемый гиперлексемный ряд является регулярное появление соответствующего лексического компонента ("лишиться того, что названо мотивирующим существительным") в серии производных с аналогичными аффиксами (ср.: деньги - обезденежеть, рыба - обезрыбеть, лес - обезлесеть и т.д.).

Регулярное появление определенных лексических компонентов в связи с присоединением к производящей основе тех или иных словообразовательных аффиксов наблюдается и в парах слов вода (водный) - надводный, вода (водный) - подводный, водяной - водянистый. Ср.: вода:надводный = бровь:надбровный = класс:надклассовый ("находящийся выше (в прямом или переносном смысле) того, что названо словом, которым в свою очередь мотивировано суффиксальное мотивирующее прилагательное"); вода:подводный = кожа:подкожный = облако:подоблачный ("находящийся ниже того, что названо словом, которым в свою очередь мотивировано суффиксальное мотивирующее прилагательное"); водяной:водянистый = болотный:болотистый = лесной:лесистый ("обладающий (в большом количестве) тем, что названо мотивирующим словом").

Таким образом, ряд перечисленных выше словообразовательно связанных слов образует одну гиперлексему, т.к. каждое слово в этом ряду принадлежит к таким относительно регулярным структурно-семантическим моделям, в которых применение той или иной формальной операции к производящей единице сопровождается предсказуемым (имеющим аналоги) семантическим сдвигом в одном из лексических значений производной.

Выделение такой единицы как гиперлексема самым тесным образом связано с самой сущностью процесса коммуникации. Именно в гиперлексеме, на наш взгляд, содержится тот инвариант смысловой, точнее - лексической информации, который позволяет, используя в разных синтаксических условиях разные лексемы одной гиперлексемы, продолжать говорить об одних и тех же денотатах, но с какими-то служебными или даже лексическими, но регулярными, добавками (3).

Неосознанное использование гиперлексемы характеризует как обычную лексикографическую практику (гнездовой и отсылочный способы представления лексического состава языка в словаре), так и прикладную лексикографию, работы по построению словарей ключевых информационно-смысловых единиц текста /10/.

За основу словника словаря были взяты два списка лексических единиц из частотных зон. Это решение было продиктовано стремлением показать процедуру конструирования гиперлексем на разных участках лексического состава языка - в зоне высокочастотной лексики (f >= 500) и в зоне среднечастотной лексики (f = 10). Выборка производилась из "Частотного словаря русского языка" Л.Н. Засориной.

При сравнительном рассмотрении исходных лексемных списков двух частотных диапазонов можно обнаружить, что они различаются между собой по ряду параметров.

Существенное различие между ними проявляется прежде всего в характере распределения лексем по частям речи. В списке слов с f >= 500 (всего 202 слова) более или менее равномерно представлены 9 частей речи - имена существительные, имена прилагательные, глаголы, наречия, местоимения, числительные, предлоги, частицы. Доля служебных слов в этом списке составляет 26 %. В списке слов с f = 10 (всего 119 слов) доля служебных слов составляет всего 3 %, а среди знаменательных частей речи доминируют имена существительные.

Сравниваемые исходные списки различаются между собой и по семантическому объему входящих в них лексем. Известно, что между частотностью слова и его семантическим объемом (количеством словарных значений) существует определенная зависимость - обе эти квантитативные характеристики находятся в отношении прямой пропорции: чем больше значений у слова, тем выше его частота (4). Лексемы первого частотного диапазона отличаются более высокой полисемией, среднее количество значений у одного слова здесь равно 6,2. Лексемы с f= 10 характеризуются сравнительно низкой полисемией, в этой частотной зоне среднее количество значений у одного слова равно 1,9.

Важно отметить, что по опытным данным наблюдается статистическая зависимость между частотностью производных слов и их словообразовательным потенциалом, т.е. способностью слова быть производящей основой /1/, /11/. Наиболее частотные слова обладают в среднем и большим словообразовательным потенциалом. Однако, следует отметить, что среди высокочастотной лексики выделяется особая группа слов, словопорождающая способность которых равна нулю. Это относится к служебным и некоторым так называемым полуслужебным словам, к последним мы причисляем местоимения. Среди слов с f >= 500 самым мощным словообразовательным потенциалом обладают знаменательные слова. По нашим данным одно непроизводное знаменательное слово имет 37,5 аффиксальных дериватов. Словообразовательный потенциал местоимений и служебных слов приблизительно одинаков: у местоимений он равен 4,4, у служебных слов - 3,1. Таким образом, можно утверждать, что словообразовательная активность слов зависит не только от частоты употребления слова, но и от того, к какой части речи оно принадлежит. Слова с f = 10 обладают значительно низким словообразовательным потенциалом - в среднем на одно непроизводное слово приходится 10,7 дериватов.

Таким образом, источником для формирования гиперлексем послужили все дериваты (за исключением сложных слов), входящие в словообразовательные гнезда, возглавляемые лексемами из двух исходных списков. Всего для анализа было привлечено приблизительно 7240 лексем, сведенных в 949 гиперлексем.

На основе лексем из исходного списка слов с f >= 500 сформировано 734 гиперлексемы. Наибольшее количество гиперлексем, как и ожидалось, получено от знаменательных слов - 633 гиперлексемы, от служебных слов и местоимений получено соответственно 67 и 34 гиперлексемы. На основе лексем из исходного списка слов с f = 10 сформировано 215 гиперлексем.

Разница между исходными списками лексем из различных частотных диапазонов как бы проецируется и на уровень полученных результатов. Квантитативное исследование материала показывает, что гиперлексемы, полученные в зоне высокочастотной лексики, отличаются от гиперлексем, полученных в зоне среднечастотной лексики, как по размерам (по количеству лексемных составляющих), так и по семантическому объему.

Среднее количество лексем в одной гиперлексеме из высокочастотной зоны равно 6,7. Причем, в гиперлексемах, сформированных на основе знаенательных слов, в среднем насчитывается 7,4 лексем, а в гиперлексемах, полученных от служебных слов и местоимений, в среднем насчитывается соответственно 2,5 и 3,2 лексем. В зоне среднечастотной лексики одна гиперлексема в среднем насчитывает 5,9 лексем.

Прежде чем перейти к количественной оценке семантического объема гиперлексемы, поясним, что подразумевается под этим семантическим объемом.

Семантический объем гиперлексемы представляет собой сумму семантических объемов составляющих ее лексем. Но эта сумма не является результатом простого, механического сложения значений лексем. Каждая из лексем, включаемых в гиперлексему, обязательно имеет лексико-семантическую соотносительность или тождественность по одному или более значений с одной или несколькими лексемами в составе гиперлексемы. В таком случае роста семантического объема гиперлексемы не происходит, а просто фиксируется информация о том, по какому из значений совокупного семантического объема гиперлексемы соотносятся или имеют тождество лексемы, объединяемые в ту или иную гиперлексему. Но в то же время в наборе лексических значений любой из лексем могут обнаруживаться одно или более индивидуальных, только ей присущих значений, за счет которых и происходит прирост семантического объема гиперлексемы в целом.

Гиперлексемы, полученные на основе производящих слов из зоны высокочастотной лексики, характеризуются в целом более высокой полисемией. Среднее количество значений у одной гиперлексемы здесь равно 5,2. Несмотря на то, что общее количество значений в семантической структуре гиперлексем, сформированных на основе знаменательных слов, более чем в 5 раз превышает общее количество значений, имеющихся в семантической структуре гиперлексем, сформированных на базе служебных слов, в среднем более полисемичными оказываются гиперлексемы именно у служебных слов, чем у знаменательных (ср.: среднее количество значений у одной гиперлексемы для знаменательных слов равно 4,9, а для служебных - 7,9). Столь высокая полисемия "служебных" гиперлексем объясняется, по-видимому, особенностями семантической структуры служебной лексики, качественно отличающейся в этом отношении от знаменательной лексики. Среднее количество значений в одной гиперлексеме, полученной в зоне среднечастотной лексики, равно 2,9.

Словарная статья в словаре гиперлексем имеет форму матрицы. В левой части матрицы располагаются лексемы, образующие гиперлексему, под каждой из которых в вертикальном столбце цифрами обозначены ее значения в той последовательности, в какой они представлены в ССРЛЯ. В правой части содержится информация о сводной семантической структуре гиперлексемы. Знак "+" на пересечении оси значений лексемы и оси значений гиперлексемы показывает, по каким значениям осуществляется лексико-семантическая связь между элементами гиперлексемы.

Тексты толкований значений каждого из слов, входящих в данную гиперлексему, даются отдельными списками. Отсылка к нужным толкованиям осуществляется за счет совпадения номеров значений лексем в тексте толкований значений и в схеме словарной статьи гиперлексемы.

Покажем для примера схему словарной статьи гиперлексемы народ.

В предлагаемом способе словарного представления лексического материала прозрачно просматриваются не только нити, связывающие семантические объемы лексем с совокупным семантическим объемом гиперлексемы, но и лексико-семантические связи между отдельными лексемами в составе гиперлексемы. Во-первых, мы имеем информацию о словообразовательной активности того или иного значения. В гиперлексеме народ, например, наиболее активным значением является 1-ое значение, которое реализуется во всех лексемах. Менее активны 3-е, 2-ое и 4-ое значения (реализуются в двух лексемах), словообразовательный потенциал остальных пяти значений равен нулю. Во-вторых, здесь наглядно представлена картина лексико-семантической соотносительности производных и их производящих, показано соотношение их семантических объемов. Так, например, семантический объем производного народный включает все значения своего производящего народ и, кроме того, развивает свои самостоятельные значения. Семантическая структура производного народность также строится на базе одного значения производящего народный и на основе развития самостоятельных значений. И, наконец, мы можем судить о количестве новых значений у дериватов, о степени усложненности их семантической структуры.

Квантитативно-системное исследование словаря гиперлексем показало, что гиперлексема, как и любая другая единица лексической подсистемы языка, обладает весьма существенными квантитативными свойствами.

Объектом квантификации является каждая отдельная гиперлексема, которой были приписаны следующие количественные характеристики - частота, семантический объем, лексемный объем, категориальный объем.

Частота определялась путем суммирования частот входящих в нее лексем. Информация о семантическом и лексемном объемах гиперлексемы в эксплицитном виде представлена в словарной статье. Так, в гиперлексему народ входит 7 лексем, а семантический объем ее равен 10. Категориальный объем характеризует способность лексемы сочетать в себе свойства нескольких частей речи одновременно. Например, в гиперлексеме народ имеются представители двух частей речи, следовательно, категориальный объем этой лексеамы равен 2.

В результате поэтапного системно-квантитативного анализа было установлено, что зависимость частоты гиперлексемы от ее семантического объема близка к прямой пропорции - с ростом частоты гиперлексемы возрастает ее семантический объем. Причем, гиперлексемы, сформированные на базе служебных и полуслужебных частей речи, имеют частоту, значительно превышающую частоту гиперлексем, сформированных на базе знаменательных лексем. Так, например, самая низкая частота для гиперлексем, сформированных на базе служебных и полуслужебных слов равна соответственно 653,0 и 859,7, а для гиперлексем, сформированных на базе знаменательных лексем из разных частотных диапазонов, она равна 15,6 (f >= 500) и 6,9 (f = 10).

Семантический объем гиперлексемы существенно зависит от ее лексемного объема. Ведь каждая из лексем, как правило, вносит свой вклад в совокупную семантическую структуру гиперлексемы. Причем, гиперлексемы с количеством значений от 1-го до 4-х, сформированные на базе знаменательных слов из более высокого частотного диапазона, отличаются меньшим лексемным объемом, чем гиперлексемы с таким же количеством значений, сформированные на базе среднечастотных слов. Для гиперлексем, полученных от служебных и полуслужебных слов, эта зависимость имеет много отклонений и не является столь выраженной. Это связано прежде всего с тем, что служебная и полуслужебная лексика отличается, с одной стороны, высокой полисемией, а с другой стороны, низким словообразовательным потенциалом.

С ростом семантического объема гиперлексемы пропорционально возрастает ее категориальный объем. У гиперлексем, полученных от знаменательных слов, категориальный объем значительно выше, чем у гиперлексем, полученных от служебных и полуслужебных слов. По-видимому, это объясняется низкой словообразовательной активностью служебной и полуслужебной лексики, которая в меньшей степени способна к образованию разнообразных в частеречном отношении производных слов. Служебные слова если и образуют производные, то зачастую это производные той же части речи, правда, с добавлением какого-либо усложняющего или уточняющего элемента. Ср., например, где - где-либо, где-нибудь, где-то, нигде, кое-где, негде; за - из-за, по-за; между - промежду, меж, промеж; над по-над; когда - когда-либо, когда-нибудь, когда-то, кое-когда, никогда, некогда; мой - по-моему и др. Напротив, большинство знаменательных слов способно выступать в качестве мотивирующей базы для слов практически всех частей речи. Таким образом, можно констатировать, что с ростом числа слов разных частей речи, входящих в гиперлексему, растет ее семантический объем. Ведь каждая лексема, входящая в гиперлексему, имеет в своей семантической структуре хотя бы одно, а может и несколько, специфичных для нее, как представителя той или иной части речи, значения.

В настоящее время продолжается работа по квантитативно-системному исследованию ряда других параметров словаря гиперлексем, таких, как 1) сила семантической связи между отдельными членами гиперлексемы, 2) зависимость между типом словопроизводства производного и степенью семантической связи с производящим, 3) зависимость между типом лексического значения (основное, периферийное, прямое, переносное) и степенью его активности в организации совокупного семантического объема гиперлексемы и т.д.

Эти и некоторые другие данные будут способствовать более глубокому пониманию принципов устройства лексической подсистемы языка, а также ущественно дополнят картину лексико-семантической соотносительности производных и производящих слов в русском языке.

Примечания

(1) О "лексикографической параметризации языка" см. [5-6]

(2) Наличие изоморфизма в отношениях между формами словоизменения и между отдельными суффиксальными производными и их производящими неоднократно подчеркивали Ал. Дювернуа [4, с. 2, 3], Г.О. Винокур [3, с. 439], Л.В. Щерба [12, с. 75, 76]. (3) См. [2], [7], [8], [9].

(4) Зависимость семантического объема от частоты употребления слова впервые была сформулирована Дж. Ципфом [13].

Литература

1. Бартков Б.И. Количественные методы в дериватологии // Исследование деривационной подсистемы количественным методом. Владивосток, 1983.
2. Борода М.Г., Поликарпов А.А. Закон Ципфа-Мандельброта и единицы различных уровней организации текста // Учен. зап. Тартуского ун-та. 1984. Вып. 689.
3. Винокур Г.О. Избранные работы по русскому языку. М., 1959.
4. Дювернуа Ал. Об историческом наслоении в славянском словообразовании. М., 1867.
5. Караулов Ю.Н. Лингвистическое конструирование и тезаурус русского языка. М., 1981.
6. Караулов Ю.Н. Современное состояние и тенденции развития русской лексикографии // Советская лексикография. М., 1988.
7. Каримова Г.О. Гиперлексемная группировка слов как способ представления системности лексики // Учен. зап. Тартуского ун-та. 1989. Вып. 872.
8. Каримова Г.О., Поликарпов А.А. Принципы выделения гиперлексемы как единицы лексической системы языка // Деривационные типы и гнезда в синхронии и диахронии. Владивосток, 1989.
9. Поликарпов А.А. Логическое пространство единиц лексической подсистемы языка и квантификация соотношений между ними // Прикладная лингвистика и автоматический анализ текста. Тез. докл. Тарту, 1988.
10. Сахарный Л.В. Частотный словарь индексирования. Пермь, 1974.
11. Тулдава Ю.А. Проблемы и методы квантитативно-системного исследования лексики. Таллин, 1987.
12. Щерба Л.В. Восточно-лужицкое наречие. Пгр., 1915.
13. Zipf G.K. Human Behavior and the Principle of Least Effort. Cambridge, 1949.

An Attempt of Quantitative-Systemic Study of Russian Hyperlexemic Dictionary

Karimova G.O.

Summary:
The paper describes principles of choosing hyperlexemes as units of language lexical system. Principles of creating hyprlexemic dictionary of Russian are suggested. Some of its quantitative-systemic characteristics are present.

Towards the Principles of Segmentation of a Coherent Text

M.G.Boroda
Ruhr University at Bochum, Germany
Phone: (+49) 2323-38-98-82
Fax: (+49) 2323-38-98-11

Topical paper

AREA: Textual units and structure

Summary:
The paper discusses the segmentation of a coherent text as a general problem of quantitative-systemic text study.

Basing the considerations on the analysis of such "absolutely coherent, undividable" texts (having in their fixation no segmentation sings) as musical ones, it is demonstrated that:
1. the formation of units of different levels in a musical text is subject to general principles actual in music of at least 17th to the 20th century
2. each unit is build of units of lower level; its formation is based on the metrorhythmic relations of the neighbouring constituent units - in particular, on their relations as to the length, resp. their position in a measure
3. the basic relations and general principles governing segmentation of musical texts prove to be actual for texts written in natural language - especially for poetic texts, were they regulate the building of verse lines.

Specifically, the study showed that the verse lines, or (in rather exclusive cases where the verse line is clearly divided into [as a rule,] two parts) their parts are build by an "attraction" of a word to the following longer word (that is to say, to of a word consisting of more syllables) and the "attraction" of a word to the previous more stressed word, and some other tendencies which are based on the relations of the neighbouing word as to their length in syllables, resp. their grade of metric accentuation.

Similar forces prove to decide upon the isolation of basic (motif und mictomotif like-) units in music (in a melody). Specifically, these units are isolated by (a) a tendency of a tone to the next tone larger in duration, and (b) a tendency of a tone to the previous metrically stronger tone.
4. the frequency of occurence of rhythmic structures of verse lines proves to be similar for lines with different metric structure and analogous to such distribution for musical texts.

The paper suggests a general method for studying segmentation processes in a coherent text, based on the classification of the relations of its elementary units as to the length and the degree of accent.

О принципах сегментации связного текста

Борода М.Г.

Резюме:
В работе обсуждается сегментация связного текста как общая проблема его квантитативно-системного изучения.

Synergetic Generative Grammar

George Silnitsky
Smolensk State Pedagogical Institute
Russia, 214000 Smolensk, Przhevalsky st., 4
Phone: (081-22)-37700

Topical paper

AREA: Synergetics and Linguistics

Summary:
Problems of applying concepts of synergetics to linguistics are discussed.

1. One of the main problems of synergetics consists in an exposition of the factors determining the genesis of complex structures of a higher (derived) level from simpler units of a lower (initial) level. We have here the nontrivial problem of the origin of qualitative novelty, irreducible to a sum of the qualities of the constituent elements; in philosophical terms, it is the perennial problem of predeterminism vs chance, freedom vs necessity, spontanity vs predictability, emergent creation vs mechanistic reductionism.

We shall distinguish between three main types of interlevel synergetic (generative) relations determined by the type of the derived unit.

1) "Energic" generative relations are connected with various transformations of physical energy resulting in a genesis of new material sructures - physical, chemical or organic (biological).

2) "Informational" generative relations represent various transformations of information in the human (or animal) psyche resulting in a genesis of semiotic (communicative) structures serving as bearers and conveyors of information.

3) "Social" generative relations constitute a combination of the first two types of relations resulting in a genesis of new social structures and human interrelations.

Independently of these three types, we shall differentiate between i n t e n t i o n a l and n o n i n t e n t i o n a l generative relations (respectively determined or undetermined by individual human volition).

The present paper is dedicated to a consideration of some synergetic aspects of the genesis of speech structures (sentences) and thus pertains to the sphere of intentional informational generative relations.

2. Speech-generating activity is synergetic in its essence. This is manifested in two main ways: in the syncretic fusion of several meanings and functions of one and the same form and in the mode of formation of complex units from simpler ones.

Synergetic syncretism may be illustrated by the concomitance of several grammatical meanings of a flection or of several multilevel functions of a base. In the latter case bases, besides expressing lexical meanings, convey different types of grammatical information - morphological (every base pertains to a certain part of speech) and syntactic (every base is characterized by a set of valencies, i.e. a system of governing functions).

On another level of investigation the synergetic approach throws new light upon the nature and volume of human language competence, in particular upon the mechanism which allows the average speaker to hold in his language memory an incredible amount of word-forms, reaching in some languages (such as Basque or Archinsk) the astronomical number of hundreds of thousands per verbal base. The synergetic viewpoint is that wordforms are initially represented in the individual language competence not as "ready-made" units automatically reproduced in speech, but in the dismantled form of disparate bases and morphological paradigms, stored in the memory as two separate lists and operationally assembled in the very process of speech.

But the main manifestation of the synergetic nature of language is to be seen on the syntactic level, in the process of sentence generation.

3. This process has in the last decades constituted the subject matter of N.Chomsky's transformational generative grammar. Sentence generation is modelled here as a series of formal transformations of a constant, apriori given content - the "basic", or "deep" structure.

The goal of synergetic generative grammar, in distinction to its transformational counterpart, is an elaboration of an emergent model of sentence generation where the sentence is treated as a qualitatively new unit of a higher level, irreduceble to any initial complex unit.

The notion of valency plays a cricial role in the synergetic model. The generative role of valency is determined by the fact, that lexical bases (lexemes) synergetically combine two aspects: substantial and relational. Substantional characteristics (lexical: meaning of the lexeme; grammatical: its morphological form) refer to the lexeme itself. The relational features of a lexeme characterize it functionally, through its relations, grammatical and semantic, with other lexemes in the structure of the sentence. Fron this point of view, every lexeme "inducts" two parallel, mutually independent sets of "positions" ("parts of the sentence"), syntactic and semantic, dyadically interrelated with one another. Each "vertical" correlation of a certain syntactic with a certain semantic position, inducted by a lexeme, will be termed a "functionally combined", or "bilevel" position (further referred to simply as "position").

The connection between lexemes in a sentence is determined by the fact that the positions inducted by a lexeme are "filled in" by other lexemes, whereas the first lexeme, in its turn, constitutes the substantial content of a position, inducted by some other lexeme. We shall qualify the valency of the first lexeme as "descendent" in the first case and "ascendent" in the second.

4. The set of lexemes stored in the language memory of a speaker and his ability to arrange them into sentences constitute his "language competence" (as understood by N.Chomsky). Byt it should be noted that this language competence does not generate speech endogenously, "from within itself", but serves as a means of verbalizing a certain message ("c o m m u n i c a t i v e c o n t e n t") determined by extralinguistic factors.

Thus, in the framework of the synergetic model, in distinction to the transformational approach, the genesis of the sentence is determined not by a single generative system, but by the interaction of two initially autonomous informational complexes - the communicative content and the language competence, each of which is characterized by its own specific set of components and principle of organization. Both these initial complexes are diffuse and multidimensional, while the sentence generated by their interaction has a unidimensional discrete structure. The synergetic process is thus connected here with a radical simplification of the two initial complexes, with an ultimate reduction of their dimrnsional parameters and a stricter syntagmatic organization of the elements selected.

The genesis of speech being an intentional process, the synergetic model of sentence generation must include a third autonomous component - the c o m m u n i c a t i v e p u r p o s e of the speaker, i.e. the motive force that brings into interaction his language competence with a certain communicative content.

5. The syntactic basis of the sentence is constituted by the p r e d i c a t i v e relation between the grammatical subject and the preicate. Traditional grammar treats both these syntactic positions as the "main parts of the sentence", whereas in L.Tesniere's valential syntax it is only the (verbal) predicate that is regarded as the "nucleus" of the sentence, while the subject is considered as an "object in the nominative case".

It should be noted, however, that the predicative relation is characterized by a bilevel structure, a synergetic junction of two discrete syntactic relations: the predicate not only "governs" the nominative case of the subject but likewise agrees with it in number and gender (in languages where these morphological categories are represented).

Besides this double syntactic function, the subject has some other characteristics which bespeak its specific syntactic status in the sentence:

1) The subject is the most frequent, and in some languages (e.g. English) - the only obligatory element of the predicative environment (in the indicative mood).

2) The subject is the only part of the sentence that is always directly connected with the predicate and therefore serves as its "syntactic marker"..

3) The nominal subject is always represented in the nominative case (and is thus "morphologically marked").

These characteristics of the subject determine its specific communicative function. As stated above, the genesis of a sentence consists in a transformation of a certain multidimensional communicative content into unidimensional syntagmatic structure. Any linear sequence has a certain beginning. The main function of the subject is precisely that, independent of its meaning, it constitutes such a "starting point" in the construction of the sentence.

From the formal point of view, the fixed morphological form of this syntactic position, independent (in distinction to all the other types of verbal complements) of the lexical type of the verb, allows the speaker to choose a certain noun as the subject at the very outset of the sentence-generating process, before the other components of the sentence (in particular, the predicate) find their explicit verbalization.

In the semantic line this initial subject highlights that element of the communicative content which presents itself to the speaker as the optimal juncture of his language competence with the informational complex undergoing the process of verbalization.

The proposed definition of the subject, in distinction to the notions of "theme", "topic" etc., is of a purely functional nature: it is based not upon semantic criteria, but upon considerations of "convenience" (from the speaker's point of view) in the deployment of the sentence on the background of the preceding and in the perspective of the following context. In principle, any element of the denotational situation may be expressed by the subject, thus fixing a certain perspective for the subsequent construction of the sentence. On this functional-communicative criterion the subject may be termed the "e x p o n e n t" of the sentence.

The second stage in the process of sentence generation consists in filling in the position of predicate with an appropriate verbal lexeme, inducted by the subject semantically (in conformity with the communicative content) and grammatically (through the syntactic relation of agreement). The fact that it is the predicate that agrees with the subject, and not vice versa, demonstrates the secondary communicative function of the former in the sentence.

On the other hand, the predicate, characterized by the most diversified and ramified descendent valency, constitutes the structural nucleus of the sentence determining its syntactic framework. On the basis of this structure-modelling function of the predicate we can designate it (together with its descendent syntactic environment) as the "e x p o n a t e" of the sentence.

Thus, the synergetic process of sentence generation consists of two discrete operations: a) the singling out from the multidimensional continuum of the communicative content of a "supporting element" ("point d'appui") of the forthcoming sentence and designating it by a noun in the nominative case, and b)the junction of this initial basic element with a semantically adequate "superstructure" with the predicate in the nuclear position. In other words, the genesis of the sentence is determined by a consecutive choice and junction of a certain exponent with an appropriate exponate.

As shown above, the exponent in its typologically constant form (nominative case) is mainly oriented on the communicative content, while the valential structure of the exponate is to a large degree determined by the specific grammatical system of this or that particular language.

It follows that the bicomponentional functional stryctureof the sentence reflects the synergetic interaction of its two generating informational complexes; our thesis that the synergetic process as such determined by an interaction of two (or more) autonomous systems thus finds its corroboration not only at the initial, but likewise at the resultant phase of sentence generation.

Синергетическая порождающая грамматика

Сильницкий Г.Г.

Резюме:
Обсуждаются проблемы приложимости концептуального аппарата синергетики к лингвистике.

Содержание Contents

От оргкомитета
From Organising Committee 5

Andreyev S. and Vislinskaya E. Establishment of Differential Characteristics of Inductive Classes by Means of Discriminant Analysis (on the Material of Poetic Works by English Romanticists) 165

Agranat T.B. The Development of Analytism in the Hungarian Declension 6

Alexeyenko L., Gryaznuchina. The Computer Version of the "Part-of-Speech" Concept 95

Anoshkina J.G. Morphological Processor for the Russian Language 7

Arold, Anne Vergleich von verschiedenen Methoden der Clusteranalyse in Linguistishen Forschungen (Cluster-analyse in der Linguistik) 8

Баскевич В.М. Статистический анализ лексического состава текстов функциональных стилей
(Baskevich V.M. Statistical Analysis of Lexical Units in Texts of Different Functional Styles) 20

Beliaeva L.N. Theoretical Principles of Linguistic Database Design in Natural Language Processing 23

Benko V. and Kostolansky E. Some Quantitative Data on the Machine-Readable Version of KSSJ 172

Boroda, Moisey G. Towards the Principles of Segmentation of a Coherent Text 227

Van den Bosch, Antal; Content, Alain; Daelemans, Walter and De Gelder, Beatrice Analysing Orthographic Depth of Different Languages Using Data-Oriented Algorithms. 26

Breiter M.A. Lengh of a Chinese Word in Relation to its other Systemic Features 166

Budzhak, Svitlana Variable Rule Analysis of "v/u" Alternation Coustrains in Canadian Ukrainian. 32

Cherneyko L.O. and Dolinsky V.A. "Sud'ba" and "rok". Semantics of Noun as an Object of Conceptual and Quantitative Analysis 49

Chizhova L.A. and Shibaeva M.V. Multicomponent Names of International Organizations as Terminological Units 51

Choudry, Amitav Models of Bilingual Measurement and their Adaptability in the Indian Context 52

Darchuk N. The Research of Computer Paradigmatics of the Verb 59

Devos, Filip; Van Gyseghem, Nanca; Vandenberghe, Ria and De Caluwe, Rita Modeling Vague Lexical Time Expressions by Means of Fuzzy Set Theory 60

Dolinsky V.A. Moscow Students' Word Associations 66

Dolinsky V.A. and Rudakov S. Associative Thesaurus of Syndromes 69

Domeij, Richard; Hollman, Joachim and Kann, Viggo Detection of Spelling Errors in Swedish not Usinga Word Listen Clair 71

Ermolenko G.V. The Problem of Measuring Linguostatistic Pecularities of Author's Speech in Fiction Texts 175

Enguehart, Chantal and Malvashe, Pierre Automatic Natural Acquisition of a Terminology 83

Godbert, Elisabeth and Pascro, Robert A Logical Representation for the Conceptual Coherence of a Sentence 89

Gryaznuchina T.A. and Alexeyenko L. Computer Model of the Suffix Zone in Ukranian 95

Hrebicek, Ludek The Stairway of Subsustems 210

Hue, Jean-Fracois Lexical Constraint Grammars 96

Hsin-Hsi, Chen and Yue-Shi Lee Approximate N-Gram Markov Model For Natural Language Generation 43

Kapitan M.E. From Latin To Mordern Romance Languages: Testing Regularities for Words System Features Evolution 176

Karapetjanc A.M. Affinity of Phonetical and Graphic Representations of the Basic Units of Chinese 179

Каримова Г.О. Опыт квантитативно-системного исследования словаря гиперлексем русского языка
(Karimova G.O. An Attempt of Quantitative-Systemic Study of Hyperlexemic Dictionary) 223

Karpilovs'ka E.A. Correlation Between Mono- and Multy-Dimensional Units Whithin Suffixal Inventory of Modern Ukrainian Language (To the problem of determination of order parameters in language system) 181

Kazakevich O.A. Minor Languages of Russia on Computer 108

Klymenko N.F. About Equilibrium in the Morphemic Subsystem of Language 184

Kolodyazhnaya L.I. The Calculation of Quantitative Characteristics of Philological Dictionaries in UNILEX-D System 109

Kolodyazhnaya L.I. and Polikarpov A.A. Study of Quantitative Correlations between Stylistics, Grammar and Polysemy of Words (On the Basis of OzhegovDictionary) 110

Kichi, Ejiri Word Frequency Distribution in Japanese Text 77

Кретов А.А. и Воронина И.Е. Лингвистическое обоснование программного синтеза слова (из материала русского языка)
(Kretov A.A., Voronina I.E. Linguistic Substantiation of Programme Word Synthesis (Using Data from Russian) 187

Krylov Y.K. Hurst's Law as a Universal Law of Quantitative Linguistics of a Coherent Text 113

Krytskaya V. Lemmatizator as the Recognizing Word Model 115

Kuang-hua, Chen and Hsin-Hsi, Chen Coprpus-Based Analyses of Adjectives: Automatic Clustering 37

Кукушкина О.В. Использование специализированной словарной БД для исследования морфологии русского языка
(Kukushkina O.V. Use of Specialised Dictionary Database for Russian Language Morphology Investigation) 116

Kuzmin L.A. Correlations Between Semantic, Derivational and Chronological Characteristics of English Adjectives 189

Ljalkova I. Systematic Characteristics of English Synonyms 191

Maslov M.U. and Garjaev P.P Fractal Presentation of Natural Language Texts and Genetic Code 193

Moscalenko T.A. Quantitative Analysis of Different Levels Lexical Units Distribution in Legislative Texts: Word forms, Lexemes, Hyperlexemes 117

Orlov Y.K. The Dynamics of Frequency Structure (Graphic Computer Pattern) 195

Orlova L. The Inflection of Nouns in the PC Aspect 119

Пиотровский Р.Г. В поисках синергетических механизмов языка
(Piotrovsky R.G. In Search of Synergetic Mechanisms of Language) 120

Покровская Е.А. База данных синонимов русского языка и ее квантитативно-системный анализ
(Pokrovskaya E.A. Database on Russian Synonyms and its Quantitative-Systemic Investigation) 121

Polikarpov A.A Evolutionary Aspects of a Language as a Natural Classification System 196

Rajnova, Dobrina Statistic Simulation of Text with Reference to Dynamics of its Units 122

Rykov V. Menzerath Law for Printed Speech 199

Saukkonen, Pauli Problems in Measuring and Interpreting Linguistic Variation 204

Savchuk S.O. Stylistic Typology of Texts on the Basis of Quantitative Analysis of Particular Sources of its Content. 201

Shaikevich A.Y. Lexicostatistical Analysis of Literary Characters (Problems and Approaches) 207

Shelov S.D. Systematic Arrangement of Terms in the Dictionary : Quantitative Approach to Linearization. 124

Silnitsky G.G. Synergetic Generative Grammar 228

Silnitskaya G. On the Problem of "Syntactic Bifurcation" (Polysemy of the English Infinitive) 205

Soldatenkova M., Sergutina S. Inter-Level Correlations of Etymological Types of English Verbs 208

Stepanov A.V. Automatic Typological Analysis of Semitic Morphology 127

Sun Da Jiang and Tsutsumi, Junya; Tomoaku, Hitta; Kotaro, Ono; Shiho,Nobesawa; Nakanishi, Masakazu An Intelligent Chinese Input System Using Statistical Information between Words 102

Sutcliffe, Richard F.E. and Bronwyn, Slater Disambiguation by Association: Two Algorithms and their Assessment 134

Sutcliffe, Richard F.E.; O'Sullivan, Donie; McElligott, Annette Creating a Large Semantic Lexicon for Nouns 140

Tsutsumi, Junya and Nitta, Tomoaki; Ono, Kotaro; Nobesawa, Shiho; Nakanishi, Masakazu Multi-Lingual Machine Translation Based on Statistical Information 147

Tuldava, Juhan Information Measures of Causality 153

Воронина И.Е., Кретов А.А. и Суворов А. Опыт автоматизации исследований русского силлаического стихосложения
(Voronina I.E., Kretov A.A., Suvorov A. An Attempt of Automatic Study of Russian Sullabo-Tonic Versification) 174

Zlatoustova L.V. Typological and Stylistic Characteristics of the Phonetic Word with Examples from some Indo-European Language. 158

Зубов А.В. Вероятностно-алгоритмическая модель порождения русского стихотворного текста
(Zubov A.B. Probabilistic-Algorithmic Model of Creating of Russian Verse Text) 161

Формат 60x90/8.Объем 29,5 п.л. Тираж 200 экз. Заказ №

В Н И И