V.B. Kassevitch, A.V. Ventsov, E.V. Yagounova

THE SIMULATION OF CONTINUOUS TEXT PERCEPTUAL SEGMENTATION: A MODEL FOR AUTOMATIC SEGMENTATION OF WRITTEN TEXT[*]

A computer model for segmenting continuous printed text into words is proposed. Texts lacking all spaces between words are analyzed using the simplest version of the Cohort Model, where the 9-syllable string of the text triggers an activation of lexicon entries that are consistent with the string's onset. As the program proceeds with analyzing the string symbol by symbol, activation is shifted away from lexical entries that are not consistent with it, and the cohort is resolved when a single candidate remains. The cohort's resolution naturally leads to placement of the word boundary. An important amendment to the Cohort Model algorithm is the systematic use of prosodic (accentual) information. In one of the algorithm modifications, the program does not initiate an activation of the cohort entries until the accented syllable has been reached in the left-to-right scanning of the buffer string, with all the candidate words that fail to match such a word-initial sub-string being ignored.

The model is more than 98 per cent effective in segmenting texts into words, if the texts are included in the corpus that has served as the basis for compiling the lexicon addressed by the program. No new word can be handled by the program, which suggests the issue of text segmentation as an 'autonomous' problem in its own right.

GENERAL

The text segmentation referred to in the title of this paper is thought of as an on-line operation whose task is to break down the continuous text, which lacks spaces between letters, into word-sized chunks.

The current models of speech perception differ in attaching relevance to the problem of speech segmentation. To cite just a few authors of speech perception models and theories, Anne Cutler makes it a point to develop special procedures which could detect word onsets in English by relying on the predominant tendency of English full words to have strong (unreduced) syllables as their onsets (cf. *Cutler, Norris 1988*), while others developing connectionist and conceptually similar models (such as TRACE or Shortlist) tend to look upon the word segmentation as a natural by-product of word identification procedures: it is quite understandable that as soon as the word has been identified, the problem of its boundaries loses whatever relevance it has (cf. *Frauenfehlder, Peeters 1990*; *Norris 1994*; etc.).

In fact, one could generally subscribe to the theoretical stance that the segmentation problem is overshadowed by successful word identification. Yet, there seem to be at least three real situations, where segmentation has to be acknowledged as an indispensable operation in its own right, viz.: (i) the listener is likely to be successful in following a speech

[*] This research was supported by the Russian Foundation for Basic Research under Grant No. 00-06-80247.

in progress after having tuned in at a random point; to do that, he or she should be able to detect the point where one word (not necessarily identified) ends and the next one (still to be identified) begins; (ii) any word identification normally relies on lexical access, which is impossible in cases where a *new* word is to be identified; to put it differently, man is capable of isolating novel words or nonce-words, where no identification *via* lexical access is imaginable, whereas some kind of segmentation cannot possibly be avoided; (iii) in language acquisition, the child is confronted with a continuous speech flow which is to be broken down into individual words; of course, no lexical access is of any use at this stage, since no lexicon is available. This means that, in a way, the child is expected *first* to develop certain segmentation strategies *before* she starts collecting individual words to develop her lexicon.[1] It is important to observe that such early overlearnt skills, even if not practiced actively in adult life, are never lost completely (*Bernstein 1966*).

WORD CODA AND ITS PERCEPTUAL RELEVANCE

As suggested in (i) above, one possible strategy for detecting word boundaries boils down to finding out the word's final part, i.e. its coda. Logically, it seems to be absolutely immaterial whether one finds the word's onset or its coda, as both meet precisely at the boundary point, which is the only thing that really matters. Yet, seen from the procedural perspective, the word's two extremes, i.e. its onset and coda, are not functionally equal. Speech is unidimensional; its onset, naturally coinciding with the first word's onset, is "given", while it is still to be sought where the first word ends. This coda-oriented perceptual drive attaches special weight to the word's final part.

There are a great many facts that demonstrate the special saliency of the word's coda. To begin with, word-final accent seems to be much more widespread in the languages of the world as compared to word-initial accent. The word's coda is manifestly more active in attracting accent in that *three* types of word-final accent (i.e. final, penultimate, and antepenultimate) are attested, while only *one* type of accent associated with the onset is known.

Looked at from a loosely defined Optimality Theory perspective, the 'good' structures seem to be those with fixed accent, while among the fixed-accent structures, final-accent rhythmic patterns are preferred. This is only natural if one admits that, from a perceptual point of view, lexical accent performs first and foremost a function of segmenting speech into words. Free accent, as in Russian, provides one with information about the *number* of words in an utterance, while fixed accent, like that in Polish or Czech, makes it possible to detect the word boundaries. The word-final type is more advantageous, if, under speech perception conditions, one is supposed to continuously scan the text for the word-coda points (*Касевич et al. 1990*). Numerous experimental data testifying to the special salience of the word coda and its special role in speech perception can be found in the publications of Sieb Nooteboom and his co-authors (*Nooteboom 1980; Nooteboom, van der Vlugt 1988*; etc.).

The special role of the word's coda can be seen in the regularities of language acquisition as well. In many L1 studies, young language learners have been shown to exhibit tendencies to attend to and extract stressed and final syllables, practically irrespective of the specific language type (*Echols, Newport 1992*). In other words, humans may be genetically endowed with a mechanism for detecting the word's final portions.[2]

[1] Cf.: "In order to acquire a lexicon, young children must segment speech into words, even though most words are unfamiliar to them. This is a non-trivial task because speech lacks any acoustic analog of the blank spaces between printed words" (*Brent, Cartwright 1996*: 93)

[2] According to Brent, Cartwright (op. cit.), such a mechanism relies on a capability to determine distributional regularities and phonotactic constraints. To this Saffran *et al.* (*1996*: 1926) adds: "[Our] study shows that a fundamental task of language acquisition, segmentation of words from fluent speech, can be accomplished by 8-month-old infants based solely on the statistical relationships between neighboring speech sounds. Moreover, this word segmentation was based on statistical learning from only 2 minutes of exposure, suggesting that infants have access to a powerful mechanism for the computation of statistical properties of the language input".

Morphological considerations could also be appropriate in discussing the problem of coda vs. onset relative weight. On the one hand, the root is the unrivalled semantic head of the word. It has been shown (cf. *Касевич, Яхонтов 1982*) that suffixes are definitely preferred to prefixes, which means that, in the languages of the world, root-initial structures are predominant. On the other hand, the very same fact shows that the word's coda is marked grammatically much more often than its onset. This, in turn, makes the coda more salient. In some SOV languages, even abbreviations reduce the word to its final rather than initial syllable, cf. Burmese *ni? < kuŋmjuni?* 'communist'.

It has been hypothesized (*Касевич 1998*) that the word onset vs. word-coda controversy could be reconciled, if we admit that the former functions mostly *paradigmatically*, i.e. contributing to word identification, whereas the latter does so *syntagmatically*, i.e. contributing to word segmentation. This problem needs a special discussion, which cannot be undertaken here, although some points of this paper are directly concerned with the issue above.

AN ATTEMPT AT A MORPHOLOGICALLY-BASED SEGMENTATION

If one looks at Russian-type languages with their rich inflexional morphology, it is tempting to suggest that, in languages of this type, word boundaries could be detected, relying upon information about the morphological elements associated with the word's end. Even the traditional term 'ending' seems to point to this direction, for, quite naturally, the *end* should be found where the *ending* shows up.

In reality, this means that a monitoring system is hypothesized which inserts a word boundary marker whenever it spots an 'ending'. The latter is one element from a finite set of symbols normally associated with the word's right boundary.

At least two important points are to be taken into account at this stage of our reasoning. One is concerned with the need to carry out a full phonemic analysis for at least certain 'islands' within the speech flow — normally for those sound strings that are associated with endings. How the perceptual system could pick up such strings prior to any morphological analysis remains at best obscure. The second one is, obviously, the widespread 'homonymity' of many endings with meaningless sound sub-strings embedded in word stems or any other stretches of the text. For instance, the so-called reflexive verb postfix *–s'a*, cf. e.g. *myt'-s'a* 'wash oneself', is 'homonymous' with meaningless phonemic sub-strings in such words as *s'adu* '[I] will sit down', where *s'a-* is just a phonetically embedded part of the stem *s'ad-* . In so far as the suggested mechanism would work its way blindly irrespective of any morphological or semantic information, it would break down lots of words into meaningless chunks.

Nonetheless, the problem — at least as far as the second point is concerned — appears to be empirical rather than purely theoretical: no one could guarantee that humans would never run the risk of a false segmentation to be somehow compensated at a later stage of the speech recognition process. This encouraged us to undertake an experiment where a segmentation of a Russian printed text was carried out that relied on 'endings'. The full set of the 'endings' selected to trigger word boundary placement was as shown in the Appendix below (p. 56).

It should be made clear that all the experiments described in this paper have been carried out with reference to a number of Russian printed texts (for details see below), where all the blank spaces between words have been deleted. In this way, the printed text was designed to imitate its running continuous counterpart in ordinary oral speech, which normally lacks any explicit word-boundary markers. The next stage of this experiment as it is currently planned should be an application of the same experimental design to texts transcribed in terms of standard phonemic symbols. This would be a step-by-step approximation of 'full-dress' experiments with real audible speech.

One more technical remark should be made here. All the texts used in the experiments are part of the corpus, which is the basis for a new Russian word frequency list currently in preparation. At the time these experiments were conducted, the total number of wordforms[3] that were included in the frequency list was 46,000 items. No novel utterances outside the corpus have been used in the experiments.

The experimental segmentation was carried out along the following lines: (1) at any stage of running the program, the program normally dealt with a string of 9 syllables;[4] the last (9th) syllable's symbol for the vowel was made its right boundary; (2) the word-boundary marker was inserted automatically to the right of any string of symbols that was identical to one of the 'endings' on the list; (3) the suspected words singled out by operation (2) above were matched against the lexicon entries (the list of the wordforms); (4) the string that remained as a "residue" after a successful word identification was made the initial sub-string of the 9-syllable string to be analyzed next; (5) if no 'ending' was found inside a 9-syllable string, the latter was left out as unanalyzable and then ignored.

The operations based on principles (1) through (5) were applied to a running printed text comprising 1,000 syllables, the spaces between the words having been deleted but accent marks introduced. The main results can be reduced to the following: out of 228 9-syllable strings, 22 were found to be unanalyzable. Out of 463 words actually present in the test fragment, only 55 could be identified, i.e. both segmented out *and* found to be identical to one of the lexicon entries.

In other words, even a Russian text, with its wealth of morphological markers normally concentrated at the word's right boundary, does not lend itself to a reasonably successful segmentation into words if the segmentation is based on the formal morphological features associated with the word's coda.

SEGMENTATION BY IDENTIFICATION

The next stage of our study will be concerned with what can be called 'word segmentation by identification algorithm'. We are not going to ignore the points raised above which suggest the need for certain 'autonomous' strategies leading to a segmentation of the text into words. Yet, it certainly stands to reason that the overwhelmingly predominant situation is one in which people deal with real words present in their mental lexicon. This being the case, the importance of segmentation independent of and prior to word identification is greatly diminished, since it is resorted to only in relatively marginal special cases. This makes it not only allowable but even necessary to model speech perception processes, which do not make use of any special segmenting procedures, but which, instead, achieve segmentation as a natural by-product of identification via a standard lexical access routine.

Such was the line of departure for our next battery of experiments. The test texts were 75 fragments (33,229 wordforms) excerpted from our corpus. The excerpts represented 31

[3] Here we refrain from discussing a very important problem whose crux lies in the nature of the lexicon entries, *if* the lexicon is designed for speech perception. Lev V.Ščerba and, later, Charles Hockett (*1961*) advocated discriminating in a principled way between two kinds of grammar: one for speakers and the other for hearers. It has been hypothesized elsewhere (*Венцов, Касевич 1999*) that there should be, likewise, two kinds of lexicons, with one intended to serve precisely the needs of speech perception. The crucial point of the perception-oriented lexicon lies in the fact that its entries are wordforms rather than lexemes or lemmas, for it is the wordforms that the speech perceiver directly deals with in the text to be analyzed and, finally, comprehended. The problem is too serious to be addressed here in any more detail.

[4] As a matter of fact, the program had an option to vary the number of syllables included in the buffer, in this case, the program was designed to ask the desired number, and the decision was made by the experimenter. It was observed that for the texts excerpted from periodicals, the optimal number of syllables was 12 rather than 9. It is difficult to say how such facts could be interpreted in view of the renewed polemics about man's working memory (cf. *Cowan 1999*, in press; see also: *Baddeley 1995; Baddeley, Hitch 1994; Caplan, Waters 1990; Daneman, Merikle 1996; Gathercole, Baddeley 1993; Just, Carpenter, Keller 1996; Shah, Miyake 1996, Waters, Caplan 1996; Waters, Caplan, Hildebrandt 1987*).

different authors. The lexicon, as was already indicated above, comprised 46,000 entries arranged in the order of their frequency, although, at this stage of our study, no frequency cues were made operative in the text segmentation procedures.

The computer program written for the purpose of text segmentation was based on the most elementary and 'straightforward' version of the Cohort Model (cf. *Marslen-Wilson 1987* and other publications by the same author and his collaborators; see also *Венцов, Касевич 1994* for a critical assessment). The only (yet very important) difference from the simpler version of the Cohort Model was our program's use of information about the accentual pattern of the words to be segmented out (for details see below).

It could be observed that the earlier versions of the Cohort Model, which are similar to that used in our experiments, were mostly intended for the recognition of isolated words rather than continuous speech, which means that no segmentation problem had arisen for the model (see, however, *Cole, Jakimik 1980*). In contrast, our program was explicitly designed to simultaneously achieve the two goals necessary for successful speech perception: the segmentation of the running text into words and identification of the words.

The algorithm underlying the program was based on an assumption according to which the speech information is fed into the listener's working memory in a phoneme-by-phoneme manner. From the very outset, it should be noted that the oversimplified assumption above is but a remote approximation to 'reality' as one can hypothesize it. In some important respects, it runs counter to what is already known about speech perception mechanisms even at this stage of our knowledge. First, the phoneme-by-phoneme mode of operation cannot be practiced systematically, for man's auditory mechanisms simply could not cope with the amount of information normally carried by phonemic strings because of the relatively poor resolving capacity of the auditory channel. Second, speech is at the same time linear and non-linear. It is linear in the sense that its only dimension is time. It is non-linear in the sense that any point of the speech chain can — and must — be characterised simultaneously in terms of many sources of information. Third, even if one leaves aside, as we surely do, all the real problems of the front-end acoustic (psychoacoustic) analysis, which are obviously indispensable for any full-fledged model of speech perception,[5] a 'simple' question remains: where and how does the system keep the incoming acoustic signal while performing the routine operations presupposed by the Cohort Model? The acoustic influx is constant and potentially incessant and the echoic memory is believed to be rather limited, its 'imprints' decaying fairly fast; this being the case, the working memory can be expected to 'jam up' very easily, overflowing with the unending amassing of raw acoustic material. The question above is not quite equivalent to the problem of the low resolving capacity of the auditory channel referred to a little earlier, as the Cohort Model tries to by-pass the phoneme-by-phoneme analysis (for more detail, see below), however, without escaping the effect of the 'incoming signal pressure'. To the best of our knowledge, no one has ever tried to explicitly pose this very important question, not to mention give a sensible answer to it.

Notwithstanding all the very real limitations inherent in the approach chosen, we still believe that this approach can be of interest as a way of modelling a *limit case* of speech perception. Here, perception can be construed as imitating the *reading-mode* text recognition, but lacking the opportunity to look either ahead or back in the text space, which is available in real reading.

[5] Any procedures of speech perception normally make use of the lexical access routine, i.e. at a certain point, the perceptual system matches the incoming speech chunks against their suspected lexical matches stored as the lexicon entries. This raises a very important problem sometimes overlooked in the current literature: whatever the choice, the incoming speech should be segmented into chunks of the same order and format as those present in the lexicon and, moreover, they should be described in terms of precisely those features that are applicable to the lexicon entries. Otherwise, any attempts to match the two sets of units (those "found" in incoming speech signals and those stored in the lexicon) would be absolutely senseless.

Thus, the program starts reading the first one or two symbols of the text to be segmented into words; in reality, the program works with the graphic symbols of a printed text, which will be referred to henceforth as 'phonemes'. As suggested by the simplified version of the Cohort Model, the candidate-word class (the cohort) is formed as soon as the first phonemes have been read off and entered into the memory buffer. The cohort includes all lexicon entries which begin with the identified phoneme or cluster.

As is the case with reading a foreign language text using a dictionary, the class of the candidate words becomes narrower and narrower with every subsequent phoneme taken into account. The end product of the process is just one word — *the* single-word cohort. Under real conditions of speech perception, this process of narrowing down the candidate word class is presumably greatly facilitated by filtering out low-frequency words, words that are not consistent with the listener's expectations, etc. These types of factors will take their legitimate place in the final version of the speech perception model; in our study, as mentioned above, only a limit case version is tested.

Our case is special in that the program is not made to attend to any factors concerned with grammatical and/or semantic well-formedness. Its only "concern" is to insert spaces between the strings of symbols in such a way that any resultant string could have its match in the lexicon. Consider, for instance, how the program would work with a text stretch *parkiopusteli*, that is, originally, *parki opusteli* 'there was nobody left in the parks by this time', lit. 'The parks have become empty'. The first cohort to be formed by the program would have, among other words, *pa* 'pas', *par* 'steam', *park* 'park', and *parki* 'parks' as its candidate words. If the program chooses, for instance, *pa* as the string singled out by spaces, then it fails to account for the string left, as there are no chances to exhaustively break down the remaining *rkiopusteli* into chunks that would have their lexicon matches.

Let us describe in more detail how the segmentation program works. As already indicated above, the text is copied from the respective file in a linear manner, deleting at the same time all the blank spaces. As for the punctuation marks, we chose to retain those marks that would presumably bear on the intonation pattern of the utterance; all such symbols have been merged into just one 'universal' symbol that, in this way, could show a kind of non-specific boundary. The memory buffer was filled in chunks equal to 9 syllables each. The 9-syllable string triggered the activation of the lexicon entries to form the cohort of word-candidates.

The following formal rules were made operative in forming such a cohort: (1) all the words that are to enter the cohort are to be taken from the available lexicon; (2) all the words should have one or two initial symbols in common with that (those) found as the onset of the 9-syllable string; (3) as the program proceeds with identifying the initial symbols of the 9-syllable string (strictly in the left to right manner), an on-line analysis of the accentual properties of the vowel phonemes is brought into play; as soon as the accented vowel is detected, all the cohort entries that differ in their pre-tonic features (actually in the pre-tonic syllables and in the accented syllable) from those of the 9-syllable string are no longer considered as possible candidate words; (4) the program keeps analyzing subsequent symbols within the buffer so long as at least one word within the cohort can be taken as the match for the string formed in this way; the analysis is stopped as soon as the resultant string finds no match within the cohort. Then comes another stage of analysis designed to sort out the candidate words, if they exceed just one word.

One more version of the algorithm referred to hereinafter as the "accent based algorithm" was also tested. In this version, the program does not initiate an activation of the cohort entries until the accented syllable has been reached in the left-to-right scanning of the buffer string; all candidate words that fail to match such a word initial sub-string are ignored.

This approach obviously had two effects, viz.: the number of candidate words in the cohort was reduced substantially, but the time to form the cohort was extended. The added time is not necessarily a disadvantage. Whatever the method of forming the cohort, nothing can cancel the "non-stop" character of feeding acoustic information into the perceiver's auditory mechanisms. In other words, a certain delay in starting up an analysis of the cohort is hardly dramatic, for, whatever the delay, it cannot put a stop to the automatic process of the further accumulation of acoustic information. This also means that the delay acknowledged above is effective only with reference to the very first word of the buffer string; the analysis of subsequent sub-strings is compensated, in terms of time, because of new acoustic information accumulated and made available by the time the previous sub-string has been processed. Last but not least, there are many reasons to believe that the hypothesized "lookahead" for the accented syllable is a parallel action running independently of the accumulation of information that provides the basis for identifying vowels and consonants. If this is really the case, then we can ignore the inhibitory effect of the accented syllable detection and of the subsequent procedures.

Speaking of the accent-based version of the algorithm, one more complication should be acknowledged. While forming the word-candidate classes as described above, the program should be designed to use special rules for disambiguating between unaccented or consonantal prepositions, conjunctions, particles, on the one hand, and word onsets that "sound the same", on the other. If no lexicon entry makes a good match for the suspected combination of a lexical word with a preposition (conjunction, etc.), the program "makes a U-turn" to analyze anew the onset of the buffer string.

On the whole, the accent-based algorithm makes no substantial improvement in the general reliability of the model, but it does make it work faster given a lesser number of steps in the process of the cohort resolution (and notwithstanding the delay in starting up the analysis discussed earlier). For a quantitative comparison of the different versions, see Tables 1 and 2.

RESULTS AND DISCUSSION

The main results are summarised in Tables 1 and 2 below. The tables show that the results are not dependent upon the type of the text, nor have other possible features, implicitly present in the excerpted texts (such as the difference in word length measured in letters), been observed in our data.

***Table 1.* The main results of the segmentation by identification algorithm (program based on the Cohort Model)**

Type of text	Number of authors	Number of texts	Number of the words	Largest number of steps in the cohort resolution	Aborted segmentation	Grammatically invalid segmentation
Fiction prose	4	23	7,389	11–18 (4–6)*	71 (0.96%)	30 (0.41%)
Plays	9	33	12,454	11–17 (3–5)*	136 (1.1%)	46 (0.37%)
Periodicals	18	19	13,386	15–18 (5–6)*	173 (1.3%)	44 (0.33%)
Total	31	75	33,229		380 (1.14%)	120 (0.36%)

*Figures in brackets are median values for the number of steps necessary for the cohort resolution.

***Table 2.* The main results of the segmentation by identification algorithm (program based on the Cohort Model supplemented by the use of accentual information)**

Type of text	Number of authors	Number of texts	Number of words	Largest number of steps in the cohort resolution	Aborted segmentation	Grammatically invalid segmentation
Fiction prose	14	14	10,209	9–11 (2–3)*	113 (1.1%)	45 (0.44%)
Plays	8	8	13,231	6–12 (2–3)*	144 (1.1%)	64 (0.48%)
Periodicals	9	9	9,526	11–18 (2–3)*	124 (1.3%)	31 (0.32%)
Total	31	31	32,966		381 (1.15%)	140 (0.42%)

*Figures in brackets are median values for the number of steps necessary for the cohort resolution

Let us analyze very briefly the cases where the program fails to produce an acceptable outcome. At the same time we will analyze certain theoretical implications concerned with the observed data.

The reported program chooses the simplest possible rule of reducing the set of candidates to just one word, viz. the final choice is that which comes before stopping a further symbol from being added to a sub-string of the initial 9-syllable string.

Two remarks seem to be appropriate at this point of our exposition. The first will touch upon a general problem of the relationship of algorithmic and heuristic strategies in speech perception. It has already been mentioned that no systematic use of phoneme-by-phoneme speech recognition is realistic, partly because of the low resolving capacity of the auditory channel in humans, partly because of the low quality and often degraded nature of the speech signal. This is why man's perceptual mechanisms are likely to use all sorts of heuristics, such as robust word identification relying on prosodic and other cues, and even guess-work. The heuristics of this sort are fairly time-saving, but run the risk of perceptual errors. Yet the perceptual system cannot altogether dispense with strategies which analyze speech in a time-consuming but relatively reliable way, that is, attending to the wealth of acoustic information contained in the speech signal. Such strategies are really indispensable for the recognition of less common proper names, absolutely new words, etc. (*Касевич 1983*).

The Cohort Model is halfway in relation to the two extremes; on the one hand, it begins with analyzing individual phonemes (or clusters) of the word's onsets. On the other, it stops such an analysis as soon as the decoded phonemic sub-string is found to be unique in that only one word among the possible candidates, and precisely this one, can be associated with this specific initial sub-string. For instance, if the sub-string [ˈɔrifl] has been successfully decoded, there is no need to keep analyzing the remaining sub-string [æm], since there seems to be just one English word, *oriflamme*, where such an initial sub-string can be found. (If one happens to be ignorant of this word, he or she will have to make a fuller phonetic analysis of the word.) As our program is fully consistent with this principle, it shares its 'middle-way' approach with the 'classic' Cohort Model, taking into account both (relatively infrequent) cases of fuller phonetic analysis and more parsimonious, time-saving guess-work strategies.

The second remark is concerned with certain typological issues. As we saw above, the program's final choice as regards the placement of word boundaries in practice leads to singling out *the longest possible word* of the cohort; sometimes we can observe what may be dubbed Pnonemic Garden Path. It is perhaps not absolutely self-evident, whether this is really the case with any other language, too, or whether Russian is special in this respect. If Russian *is* special, one can note a very interesting typological isomorphism between stem-level and word-level morphology: as is well known, in selecting the base-form for Russian stems, one is advised to single out *the longest possible stem* of those really attested (*Jakobson 1948*).

Once again we can hypothesize that, from the perspective of Russian , the 'good' word (stem) is the longest one.

Obviously, systematic adherence to "the longest is the best" principle may be taxing. Note one of the real examples where our program failed to produce an adequate segmentation: *muki stradanij* 'tortures of the misery' instead of the original *muk i stradanij* 'of tortures and miseries (Gen.)'. In this example, the program, in its persistent drive to find the rightmost word boundary whenever possible, has taken the conjunction *i* 'and' for the nominal plurality marker *-i*, hence the erroneous reading — for this particular context. Of course, this kind of analysis would have never been possible in the case of real oral speech or even its phonetic transcription, but we leave aside all these issues for the time being.

Some such missegmentations could be rectified with a grammatical and/or semantic analysis. For instance, in the final files produced by the program, one finds such instances as *cenna muku* 'valuable (Adj., short form, Fem., Nom. Sing.) flour' instead of *cen na muku* 'of the prices for the flour'. The string *cenna muku* (apart from its stylistic unacceptability) is impossible formally (grammatically), because it violates the rules of Adjective-Noun Agreement (Nom. in *cenn-a*, while Dat. in *muk-u*).

Following from what was said above, the program can restart the analysis if the resultant string ends in a sub-string, which is not accountable for in terms of the wordforms available in the lexicon. At present, such a restart is limited to just one attempt — the specific attempt concerned with the case of unaccented prepositions and the like (cf. p. 52 above). If this "U-turn" analysis does not yield an acceptable result, the attempt is considered aborted and the program stops functioning, giving up the lead to the experimenter. For instance, the string *v etom net nikakogo* 'there is nothing [special] about it' was segmented as *ve* (name of the letter, like *vee* in English) *to* 'that' (Nom.) *mne* 'to me' (Dat.) *tnikakogo* (no meaning at all). As the second attempt based on the "U-turn" approach did not yield any acceptable result, the program just stopped.

As can be seen from the tables, even without any 'reinforcement' from semantics or grammar, i.e. used in an absolutely straightforward formal manner,[6] our segmentation program scored exceptionally high in all the attempted tests. As reported above, 75 running texts, all in all comprising 33,229 words, have been subjected to automatic analysis based on our segmentation program. In a mere 1.14 per cent of the total cases (380), the program's work was abortive. In only 0.36 per cent of the total cases (120) the program failed to produce the segmentation that would agree with that of the original text.

We have to remember once again that no utterances, no texts outside our corpus, have been used in the experimental sessions. This means that the program has never confronted a word not in the lexicon. As a matter of fact, any such word, if encountered in a text to be segmented, would stop the program. This is obviously a very serious limitation that makes the program an insufficiently effective simulator of its natural prototype. Unlike the program, a human perceiver normally has little difficulty, if any, in picking an unknown word (or a quasi-word) out of the text and in further manipulating it. This brings us again back to the issue of segmentation as an 'autonomous' operation in its own right.

References

Baddeley A.. Working Memory. — M. S. Gazzaniga (ed.) The Cognitive Neurosciences. Cambridge, Mass., 1995, p.755–764.

Baddeley A.D., Hitch G.J. Developments in the concept of working memory. — *Neuropsychology*. 1994, **8**, N 4, *485–493*.

Bernstein N.A. The Coordination and Regulation of Movements. London, 1967.

Brent M.R., Cartwright T.A. Distributional regularity and phonotactic constraints are useful for segmentation. — *Cognition*. 1996, **61**, N 1–2, *93–125*.

Caplan D., Waters G.S. Short-term memory and language comprehension: A critical review of the neuropsychological literature. — G. Vallar, T. Shallice (eds.). Neuropsychological Impairments of Short-Term Memory. Cambridge, 1990, p. 337–389

Cole R.A., Jakimik J. A model of speech perception. — R.A.Cole (ed.). Perception and Production of Fluent Speech. Hillsdale, NJ. 1980, p. 133–163.

Cowan N. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. — *Behavioral and Brain Sciences*, **24** (in press).

Cutler A., Norris D. The role of strong syllables in segmentation for lexical access. — *Journal of Experimental Psychology*: Human perception and performance. 1988, **14**, N 1, *113–121*.

Daneman M., Merikle P.M. Working Memory and Language Comprehension: A Meta-Analysis. — *Psychonomic Bulletin and Review*. 1996, **3**, *422–433*.

Echols C.H., Newport E.L The role of stress and position in determining first words. — *Language Acquisition*. 1992, **2**, N 3, *189–220*.

[6] For a similar, yet "more sophisticated" approach see *Johnson, Pugh 1994*.

Frauenfelder U.H., Peeters G. Lexical segmentation in TRACE: An exercise in simulation. — G.T.M.Altmann (ed.) Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspective. Cambridge, Mass.; London, 1990, p. 50–86.

Gathercole S.E., Baddeley A.D. Working Memory and Language. Hillsdale, J, 1993.

Hockett Ch.F. Grammar for the Hearer. — Structure of Language and its Mathematical Aspects. Proceedings of Symposia in Applied Mathematics. 1961, **12**, *220–236*.

Jakobson R. Russian conjugation. — *Word*. 1948,.**4**, N 3, *155–164*.

Johnson N.F., Pugh K.R. A cohort model of visual word recognition. - *Cognitive Psychology*. 1994, **26**, N 3, *240–346*.

Just M.A., Carpenter P.A., Keller T. Working Memory: New Frontiers of Evidence — *Psychological Review*. 1996, **103**, N 4, *773–780*.

Marslen-Wilson W.D. Functional parallelism in spoken word recognition. — *Cognition*. 1987, **25**, N 1, *262–275*.

Norris D. Shortlist: A connectionist model of continuous speech recognition. — *Cognition*. 1994, **52**, N 3, *189–234*.

Nooteboom S.G. Lexical retrieval from fragments of spoken words: Beginnings vs endings. — *Journal of Phonetics*. 1981, **9**, *407–424*.

Nooteboom S.G., van der Vlugt M.J. A search for a word-beginning superiority effect. — *Journal of the Acoustical Society of America*. 1988, **84**, N6, *2018–2032*.

Saffran J.R., Aslin R.N., Newport E.L. Statistical learning by 8-month-old infants. — *Science*. 1996, **274**, N 5294. *1926–1928*.

Shah P., Miyake A. The separability of working memory resources for spatial thinking and language processing: An individual differences approach. — *Journal of Experimental Psychology: General*. 1996, **125**, N 1, *4–27*.

Waters G.S., Caplan D. The measurement of verbal working memory capacity and its relation to reading comprehension. —*The Quarterly Journal of Experimental Psychology*.A. 1996, **49**, N 1, *51–79*.

Waters G.S., Caplan D., Hildebrandt N. Working memory and written sentence comprehension. — *Attention and Performance XII*: The Psychology of Reading. M. Coltheart (ed.). London: Lawrence Erlbaum, 1987, p. 531–555.

Венцов А.В., Касевич В.Б. Словарь для модели восприятия речи. — *Вестник СПбГУ*. 1998. Вып. 3, с. 32–39 [*Ventsov A.V. Kassevitch V.B.* Lexicon in the model of speech perception. — *Bulletin of St. Petersburg University*. 1998. **3**. *32–39*].

Венцов А.В., Касевич В.Б. Проблемы восприятия речи. СПб. 1994 [*Ventsov A.V. Kassevitch V.B.* Problems in speech perception. St. Petersburg, 1994].

Касевич В.Б. Фонологические проблемы общего и восточного языкознания. М., 1983 [*Kassevitch V.B.* Phonological Problems in General and Oriental Linguistics. Moscow, 1983].

Касевич В.Б. Иерархия и функции слогов в слове: начала и концы. — *Фонетика сегодня*: Актуальные проблемы и университетское преподавание: Тезисы докладов. М., 1998, с. 50–52. [*Kassevitch V.B.* Hierarchy and function of syllables within the word: Beginnings vs. endings. — *Phonetics Today*: Actual Problems and University Education. Moscow, 1998, p. 50–52].

Касевич В.Б., Шабельникова Е.М., Рыбин В.В. Ударение и тон в языке и речевой деятельности. Л., 1990 [*Kassevitch V.B., Shabelnikova E.M., Rybin V.V.* Accent and tone in language and speech. Leningrad, 1990].

Касевич В.Б., Яхонтов С.Е (ред.). Квантитативная типология языков Азии и Африки. Л., 1982 [Kassevitch V.B., Yakhontov S. E. (eds.) A Quantitative typology of the languages of Asia and Africa. Leningrad, 1982].

Appendix

List of the 'endings'

t'
oj
o+j
om
o+m
ije
ije
i+je
yj
jet
ogo
o+go
aja
a+ja
yje
y+je
s'a
s'a+
yx
y+x
ij
ov
o+v
oje
o+je
ix
i+x
jem
je+m
uju
u+ju
ym
y+m
jej
je+j

is'
i+s'
jets'a
t's'a
jut
ju+t
mi
ami
a+mi
it
i+t
im
i+m
as'
a+s'
ax
a+x
os'
o+s'
jesh'
je+sh'
jeje
je+je
jo+m
jo+t
omu
o+mu
ju
ju+
ymi
y+mi
jego
am
a+m
jat

ja+t[*]
its'a
i+ts'a
jet'e
juts'a
ju+ts'a
ut
u+t
ch'
jax
ja+x
imi
i+mi
ish'
i+sh'
jev
jami
ja+mi
jam
ja+m
jemu
jus'
ju+s'
jo+ts'a
jo+sh'
it'e
i+t'e
jats'a
ja+ts'a
t'es'
ixs'a
jaja
jeshs'a
at
a+t

jo+t'e
jejs'a
je+js'a
ijs'a
i+js'a
juju
ims'a
i+ms'a
ijes'a
jegos'a
jems'a
ajas'a
jet'es'
ish's'a
i+sh's'a
uts'a
u+ts'a
ats'a
a+ts'a
us'
u+s'
joshs'a
jo+j
jo+ms'a
imis'a
oju
o+ju
ujus'a
jejes'a
jemus'a
it'es'
i+t'es'
jo+t'es'
ch's'a

[*] The plus ("+") symbols stand for the accent marks for the preceding vowels.

В.Б. Касевич, А.В. Венцов, Е.В. Ягунова

Имитация процессов перцептивной сегментации непрерывного текста: модель автоматической сегментации печатного текста

В работе предлагается вариант модели (компьютерной системы), осуществляющей автоматическую сегментацию на слова беспробельной орфографической записи русского текста. Анализ базируется на основных положениях ранней версии **модели когорты.** Программа помещает в буфер текстовый фрагмент длиной 9 слогов; по первому символу фрагмента начинается процесс образования текущей когорты путем активирования соответствующих слов в словаре. В одной из модификаций программы вводится важное изменение по сравнению с "классической" моделью когорты: образование текущей когорты осуществляется не по первому символу, а по предударной части и ударному гласному анализируемой цепочки. Буфер слов-кандидатов заполняется посимвольно до тех пор, пока символы в исходном буфере совпадают хотя бы с одним словом в когорте; заполнение прекращается, когда добавление еще одного элемента создает комбинацию, не представленную в словаре; вслед за этим начинается анализ слов-кандидатов. Результатом работы программы является выбор единственного слова, что автоматически определяет словесную границу. Эффективность процедуры сегментации составляет более 98%. Все сегментированные тексты принадлежат к корпусу, на основе которого был создан используемый словарь. Включение новых слов данной программой не предусмотрено, подобная ситуация поставит вопрос о процедуре сегментации текста как об автономной (самостоятельной) проблеме, требующей своих подходов.