Josse DE KOCK

Walter BOSSAERT

Towards an automatic morphological segmentation.

Our intention is to obtain a morphological segmentation of French acceptable to a present-day speaker or listener, for a corpus provided without any previous indexing, by means of electronic calculation.

Experience tends to prove that it is possible to achieve such a segmentation by means of exclusively formal criteria. It consists in establishing and grading these criteria and formulating them mathematically. The use of the computer guarantees objectiveness up to a significant linguistic level; the computer is an instrument of research for new rules; it guarantees the control of established rules.

The method followed by us is based on the hypothesis that the linguistic performance of human memory consists in a constant segmentation or reconstruction of the signs of the linguistic code on levels which are graded and organized each in accordance with its own rules, in function of the specific capacities of the human brain, and with a certain degree of productiveness.

No segmentation is excluded beforehand. The segmentation is implemented by means of factors of association or alternation of the separate segments, according to a law of minimal economy, as well as by quantitative or statistical factors concerning the number of different segments, their frequency on each side of the proposed division, and their internal relationship.

These factors seem to apply to a large number of languages and to the majority of French forms. Certain counter-indications and some correctives proper to the French language must be observed.

Finally it may be stated that a segmentation is not implemented in function of a scale of absolute values, but in function of the specific morphological tension of each word in relation to the single words resembling it from the point of view of morphology.

The corpus used is made up of 8,782 phonetic forms (isolated, conjugated and declined), which are provided by the 5,044 words which according to Juilland occur most frequently. At the moment we have made a start with programming the essential factors on samples.

Ghent, May 15 1969.

Josse DE KOCK
Walter BOSSAERT

Pour une segmentation morphologique automatique

L'intention est d'obtenir une segmentation morphologique du français acceptable pour un locuteur ou un auditeur d'aujourd'hui, pour un corpus donné sans intervention préalable aucune, au moyen d'une calculatrice électronique.

L'expérience tend à prouver qu'une telle segmentation est possible au moyen de critères uniquement formels. Elle consiste à établir ces critères, à les hiérarchiser et à les formuler mathématiquement. Le passage par l'ordinateur garantit l'objectivité à un niveau linguistique déjà significatif; l'ordinateur est un instrument de recherche de nouvelles règles; il assure le contrôle des règles établies.

La méthode suivie repose sur l'hypothèse que le travail linguistique de la mémoire humaine consiste en une segmentation ou une restructuration constantes des signes du code linguistique sur des plate-formes hiérarchisées et organisées chacune selon ses règles propres, en fonction des capacités spécifiques du cerveau humain et avec une certaine rentabilité.

Aucune segmentation n'est exclue d'avance. La segmentation est opérée au moyen de facteurs d'association ou d'alternance des segments séparés, selon une loi d'économie minimale, ainsi qu'avec des facteurs quantitatifs ou statistiques portant sur le nombre de segments différents et leur fréquence de chaque côté de la coupe proposée, et de leur rapport.

Ces facteurs semblent applicables à un grand nombre de langues et à la majorité des formes françaises. Certaines contre-indications et des correctifs propres au français doivent être observés.

Ainsi une segmentation ne s'opère pas en fonction d'une échelle de valeurs absolues, mais en fonction de la tension morphologique spécifique de chaque mot vis-à-vis des seuls mots qui lui ressemblent du point de vue morphologique.

Le corpus utilisé est constitué par 8702 formes phonétiques, isolées, conjuguées et déclinées, qui sont fournies par les 3044 mots les plus fréquents selon Juilland. A ce jour la programmation des facteurs essentiels est entrée en exploitation sur échantillons.

Gand, le 14 mai 1969