The main question for a researcher studying language, and lexicon of the language in particular pertains to how to define a vocabulary unit. This proves all the more important and difficult if one deals with units consisting of more than one word. This problem is very well-known and challenging for authors and editors of monolingual dictionaries, who have to make a choice as to what to record in a dictionary, which is particularly difficult in the case of multi-word units, and how to define a multi-word unit before starting their lexicographic work.

The linguistic phenomenon in question bears different names in specialized literature: *phraseological unit*, *multi-word lexical unit*, *multi-word expression*, *collocation*, as well as other labels, which are more or less specific in scope. The sheer multitude of terms used to describe word combinations attests to the complexity and ambiguity of the phenomenon under scrutiny. In our investigation, we want to describe the lexicalisation criteria for multi-word units; in other words, we intend to pinpoint the features that make certain multi-word language constructions lexicalised (i.e. treated as language units). The majority of Polish language speakers will certainly acknowledge that *biały kruk* (‘rara avis’) is a single, inseparable unit, with a clearly defined sense and recorded in dictionaries as a separate lexeme. However, what about *drzwi wejściowe* (‘front door’)? Or the binomial name *pies domowy* (‘Canis lupus familiaris’)? In our research, we will try to identify differences between such similar-looking word combinations, focusing on noun phrases and verb phrases, as they are the most frequently and readily used (and transformed) by language users.

The study will be carried out using both Polish and English language material since we assume that there is a possibility of developing a universal (to some extent), language-independent, set of criteria designed to assess lexicality of word combinations and to develop an adequate verification procedure of their lexicality status.

To achieve this goal, we will employ the latest methods of language exploration in the field of computational linguistics and natural language processing, consisting of multi-word unit structure description in a formal language in order to perform a quantitative study - using large language corpora - of a number of features of multi-word units, such as word order permutations (ie. *drzwi wejściowe* and *wejściowe drzwi*) or discreteness (ie. *wtrącić swoje trzy grosze* and *wtrącić do rozmowy swoje trzy grosze*). Corpus-based research will become our methodological foundation for working on specific word combinations which have been described in lexico-semantic networks (wordnets) for both languages in question. These word combinations are to be verified using the procedure custom-designed to facilitate assessment of the lexicality of word combinations.

It should be noted at this point that such a procedure exists for the Polish language and is utilised by the creators of Polish wordnet (*pl*. Słowosieć). However, it contains an elaborate set of rules applicable only to two-word combinations that consist of a noun and a descriptive adjective. Therefore, in order to conduct a comparative analysis, it is imperative that we expand this procedure so that it includes other types of word combinations. The English wordnet (Princeton WordNet) does not have this kind of procedure, and multiword units were included and described therein in an intuitive, non-systematic way. On the one hand, it is a testimony to the degree of difficulty of the subject matter, but on the other hand, it presents an enormous research potential that provides a basis and justification for our study. Its results will have numerous applications in Polish linguistics, comparative linguistics (not only Polish-English, since our endeavor is to develop universal rules, at least to some extent), notably for lexicology and lexicography, as well as in the field of natural language processing.