Podstawowym pytaniem językoznawcy zajmującego się opisem leksyki jest określenie, czym jest jednostka języka. Staje się ono tym ważniejsze i trudniejsze w przypadku, gdy analizujemy jednostki składające się z więcej niż jednego wyrazu. Problem ten bardzo dobrze znają autorzy i redaktorzy słowników jednojęzycznych, którzy przed rozpoczęciem prac leksykograficznych muszą podjąć próbę precyzyjnego zdefiniowania jednostki wielowyrazowej, aby wybrać te połączenia wyrazów, które winny zostać zarejestrowane w słownikach.

Te połączenia wyrazowe, które funkcjonują w języku jak pojedyncze wyrazy, w literaturze określa się różnymi nazwami - są to m.in. *frazeologizm*, *wielowyrazowa jednostka leksykalna*, *jednostka wielosegmentowa*, *związek wielowyrazowy* i wiele innych. Już sama wielość terminów świadczy o tym, z jak skomplikowanym i niejednoznacznym zjawiskiem mamy do czynienia. W naszym badaniu stawiamy sobie za cel opis kryteriów leksykalizacji jednostek wielowyrazowych, czyli innymi słowy, takich ich cech, które sprawiają, że pewne wielowyrazowe formy językowe są uznawane za zleksykalizowane (tzn. są traktowane jako jednostki słownictwa języka i trafiają do słowników), a inne – nie. Większość osób posługujących się językiem polskim przyzna z pewnością, że *biały kruk* to jedna niepodzielna jednostka, która ma zdefiniowane znaczenie i powinna być zarejestrowana w słownikach jako oddzielne hasło. A, na przykład, *drzwi wejściowe*? Albo nazwa gatunkowa *pies domowy*? W naszych badaniach postaramy się odnaleźć różnice pomiędzy takimi podobnie wyglądającymi połączeniami wyrazów, skupiając się przy tym na frazach rzeczownikowych i czasownikowych, jako najczęstszych i najchętniej używanych (a również i przekształcanych) przez użytkowników języka.

Prace będą prowadzone jednocześnie dla języków polskiego i angielskiego, ponieważ zakładamy, że istnieje możliwość opracowania do pewnego stopnia wspólnych, niezależnych od języka, kryteriów oceny leksykalności połączeń wyrazowych.

W naszym badaniu wykorzystamy najnowsze metody eksploracji języka opracowane na gruncie językoznawstwa komputerowego oraz przetwarzania języka naturalnego, polegające na opisie struktury jednostek wielowyrazowych w języku formalnym, aby na podstawie dużych korpusów tekstów zbadać ilościowo niektóre cechy tych jednostek, np. zmienność szyku (np. *drzwi wejściowe* i *wejściowe drzwi*) czy separowalność (np. *wtrącić swoje trzy grosze* i *wtrącić do rozmowy swoje trzy grosze*). Badania oparte o korpusy staną się dla nas podstawą pracy na konkretnych połączeniach wielowyrazowych, które opisano w sieciach leksykalno-semantycznych (wordnetach) dla obu języków. Połączenia wyrazowe zostaną docelowo zweryfikowane przy pomocy specjalnie opracowanej procedury służącej do oceny leksykalności.

Należy bowiem zaznaczyć, że taka procedura istnieje dla języka polskiego i jest wykorzystywana przez twórców polskiego wordnetu (Słowosieci), jednak zawiera rozbudowany zestaw reguł dla jednego typu połączenia – dwuwyrazowego, składającego się z rzeczownika i określającego go przymiotnika. Dlatego, aby przeprowadzić analizę porównawczą, ważne jest rozszerzenie tej procedury również na inne typy połączeń, tak, by była jak najbardziej uniwersalna, tj. umożliwiała opis leksykalności również w przypadku języków innych niż polski. Warto bowiem zaznaczyć, że wordnet dla języka angielskiego (Princeton WordNet) nie posiada takiej procedury, a jednostki wielowyrazowe były doń wprowadzane i opisywane w sposób intuicyjny, niesystematyczny. Świadczy to z jednej strony o stopniu trudności tematu, jakiego się podejmujemy, a z drugiej strony niesie za sobą ogromny potencjał badawczy i daje podstawy do badań, których wyniki będą miały liczne zastosowania w językoznawstwie polskim, porównawczym, zwłaszcza w odniesieniu do leksykologii, leksykografii, a także przetwarzania języka naturalnego.