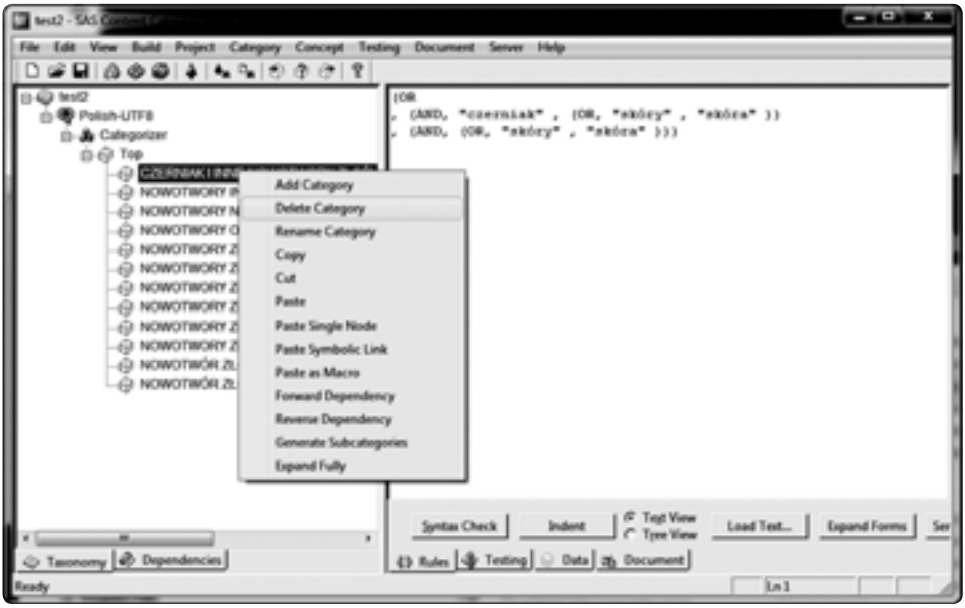


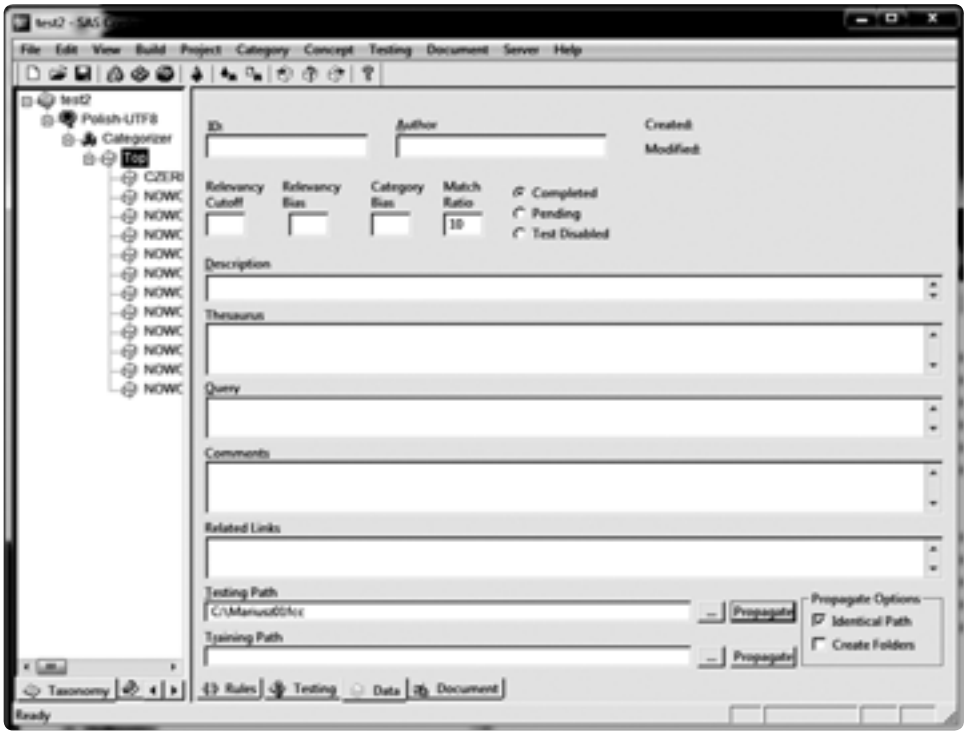
Rysunek 10.11. Menu kontekstowe istniejącej kategorii

10.1.7. Zasady używania kategoryzatora statystycznego

Jak już wspomniano wcześniej, kategoryzator statystyczny pozwala sklasyfikować w sposób automatyczny dużą liczbę dokumentów za pomocą małej liczby zdefiniowanych kategorii o szerokim zakresie. Tak zdefiniowane kategorie powinny być rozdzielne. Definicja danej kategorii jest wypracowywana automatycznie na podstawie częstości występowania terminów w dokumentach zbioru uczącego przydzielonego do danej kategorii oraz innych kategorii. Zmiana zbioru uczącego przydzielonego do danej kategorii może zatem skutkować zmianą definicji innych kategorii. Ze względu na obowiązujące zasady, którymi kieruje się kategoryzator, warto stosować następujące dobre praktyki:

1. Zdefiniować wszystkie kategorie w tworzonej taksonomii.
2. Przydzielić dokumenty ze zbioru uczącego do wszystkich zdefiniowanych kategorii. Zwykle warto wskazać 50–100 dokumentów, które są najlepszymi przykładami dla danej kategorii. Dołączone dokumenty mogą występować w różnych formatach (HTML, XML, SGML, TXT). W celu ułatwienia podpięcia dokumentów ze zbioru uczącego do poszczególnych kategorii wykorzystuje się strukturę katalogów odzwierciedlającą strukturę taksonomii. Korzeń struktury katalogów powinien być wskazany jako ścieżka do zbioru testowego dla węzła Top. Nie ma konieczności ręcznego tworzenia struktury katalogów, co byłoby uciążliwe w przypadku rozbudowanej taksonomii. Wspomniana struktura katalogów może zostać utworzona automatycznie. W tym celu wystarczy stworzyć katalog nadrzędny, a w nim katalog

podrzędny o nazwie Top. Następnie na zakładce **Data** należy wskazać katalog **Top** jako **Training Path**. Dodatkowo należy włączyć opcję **Create Folder** i nacisnąć przycisk **Propagate** (rys. 10.12).



Rysunek 10.12. Ustawianie ścieżki do zbioru uczącego dla węzła Top

Katalogi odpowiadające poszczególnych kategoriom zostaną utworzone automatycznie. Należy w nich umieścić dokumenty ze zbioru uczącego odpowiadające poszczególnym kategoriom.

3. Analogicznie postępuje się ze zbiorem testowym, przy czym należy stosować się do ogólnej zasady, że zbiory uczący i testowy powinny być rozdzielne. Zmiana nazwy kategorii w taksonomii pociąga za sobą konieczność ręcznej zmiany nazwy katalogów odpowiadających danej kategorii w strukturze katalogów zbioru uczącego i testowego.
4. Po zdefiniowaniu kategorii w taksonomii oraz przypisaniu ścieżek do zbioru uczącego i testowego następuje etap budowania reguł. W tym celu należy wybrać węzeł kategoryzatora w drzewie taksonomii, a następnie z menu kontekstowego opcję **Build->Build Statistical Categorizer**. Po zbudowaniu dany kategoryzator staje się kategoryzatorem aktywnym dla danej taksonomii. Operacje testowe zachodzą z wykorzystaniem aktywnego kategoryzatora. Automatyczne przebudowanie kategoryzatora przed wykonaniem dowolnego testu, bez konieczności wykonywania