Mariusz Piotrowski
Węzeł Centralny OŻK-SB

Analiza frekwencyjna słów używanych w dokumentach JST na przykładzie „Strategia marki i promocji miasta Augustowa w latach 2010-2015”.

*Instrukcja użytkowania programu **Antconc** dla korpusów tekstowych języka polskiego*.

Warszawa, 2014

1. Wprowadzenie.

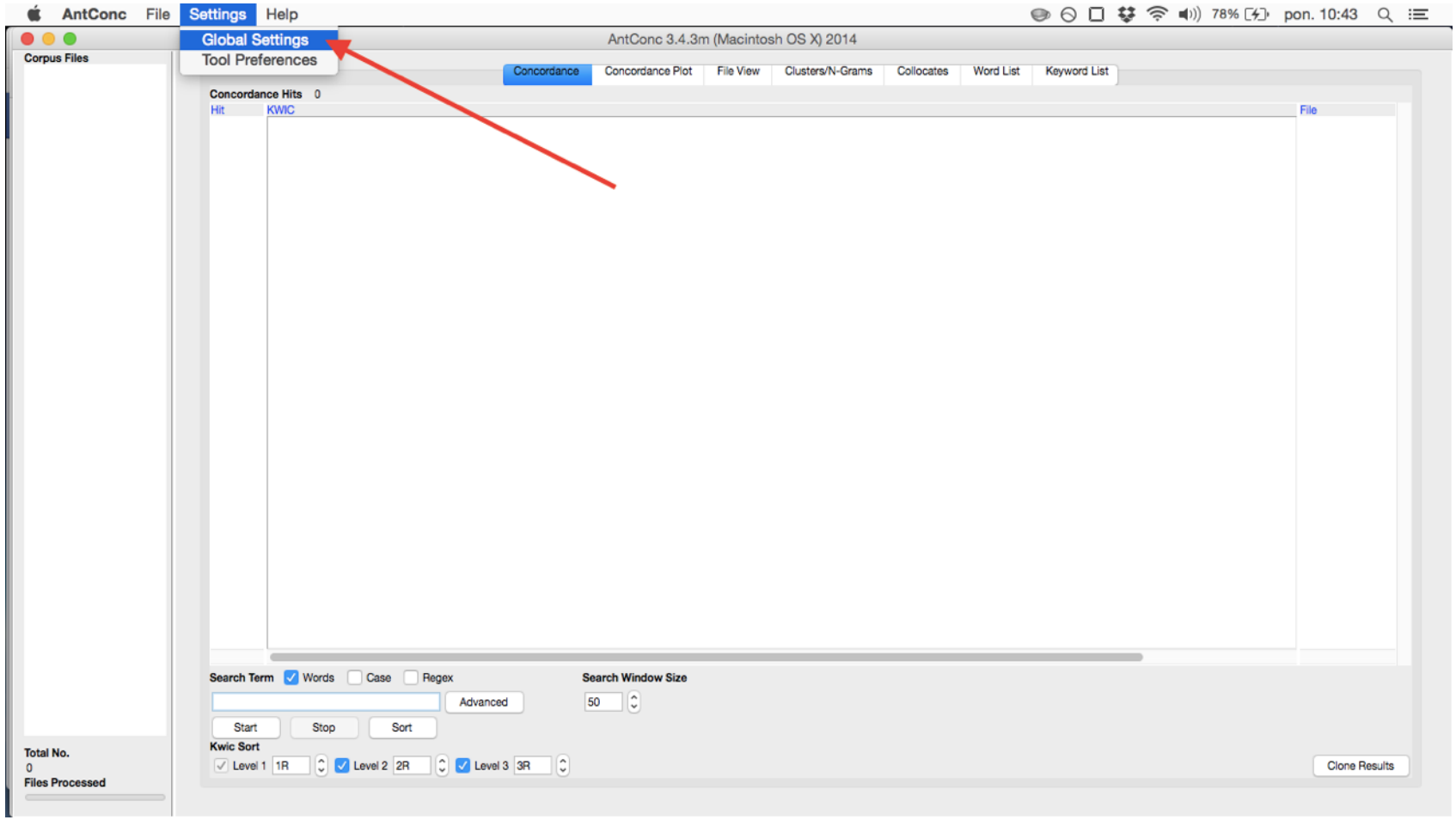
Program do analiz językowych można pobrać ze strony twórcy, pod adresem http://www.laurenceanthony.net/software.html

Działa on na najpopularniejszych systemach operacyjnych, **Microsoft Windows**, **Mac OSX** i **Linux**. Instrukcja bazuje na wersji programu 3.4.3.

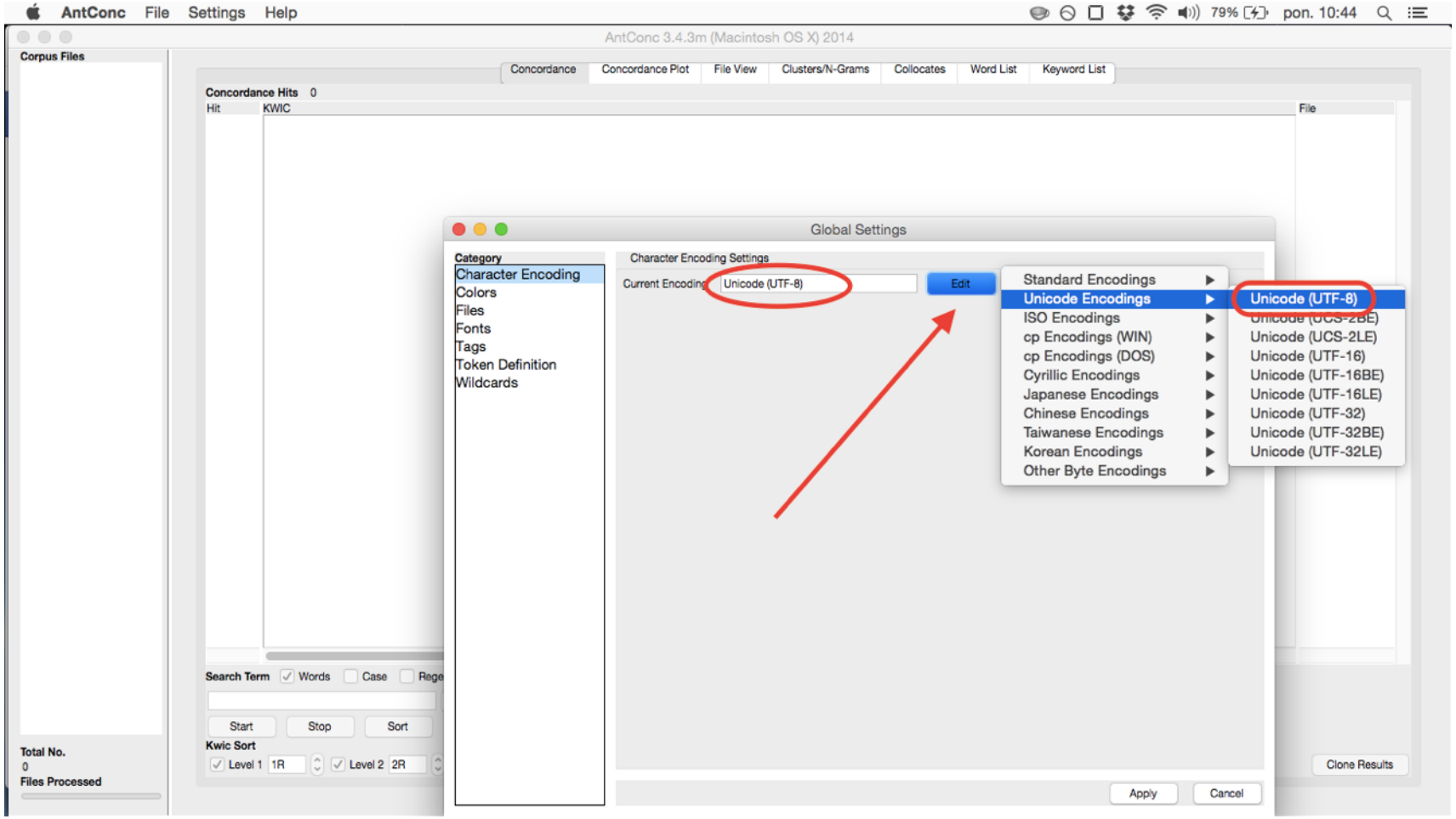
2. Przygotowanie środowiska pracy.

Korzystanie z programu **Antconc** należy rozpocząć od ustawienia parametrów środowiska pracy. Pierwszą czynnością powinno być ustawienie sposobu kodowania polskich znaków.

W menu



głównym przechodzimy do opcji **Settings**, następnie **Global Settings**.



Aby móc pracować na polskich korpusach tekstów należy zmienić w kategorii **Character Encoding - Current Encoding - Unicode (UTF-8).** (Taki sposób kodowania jest domyślny na większości współczesnych systemów operacyjnych). Jeśli w

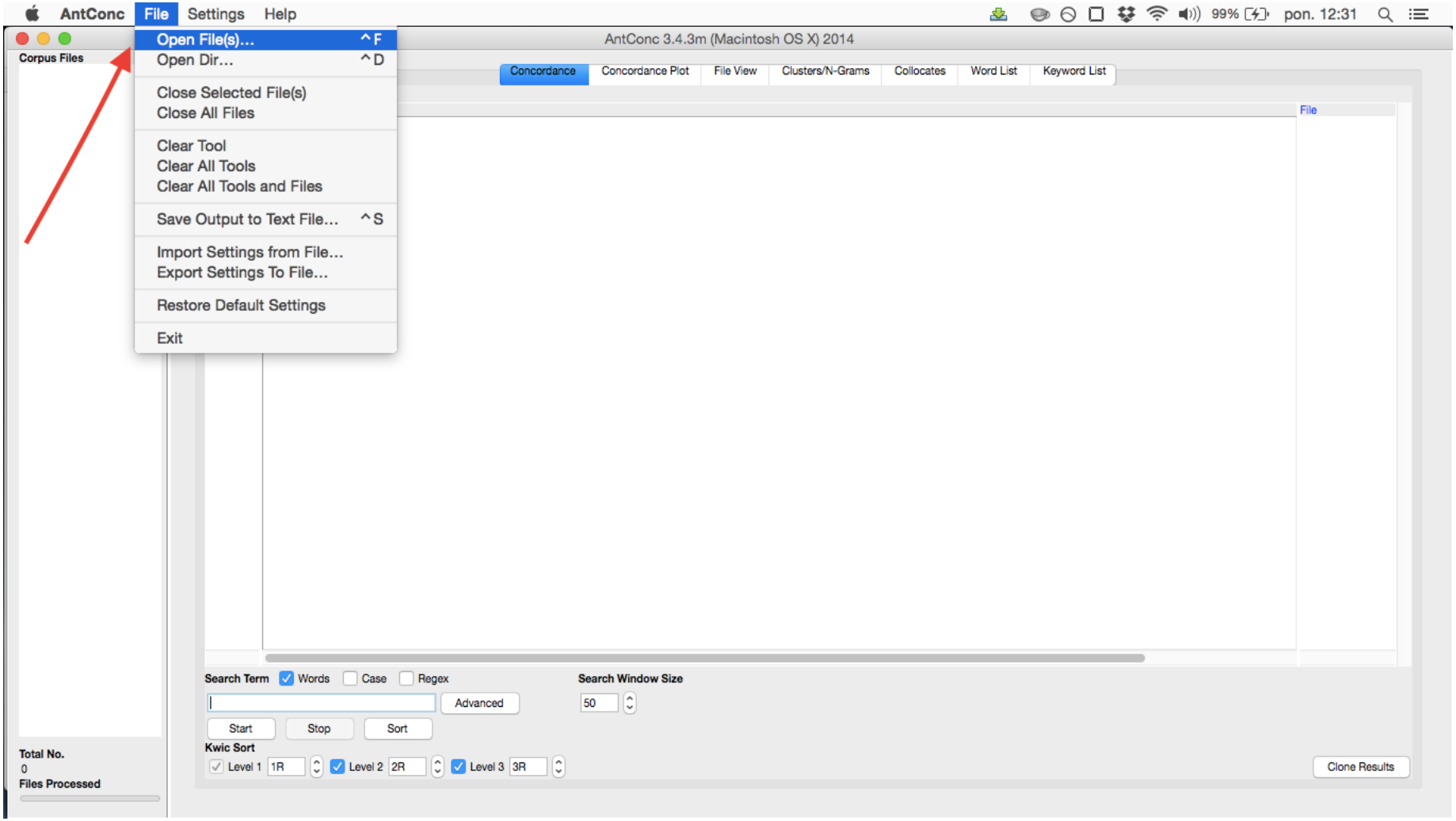
okienku **Current Encoding** jest kodowanie inne (np. CP-1250) wówczas należy wcisnąć **Edit** i wybrać **UTF-8.** Zmiany zatwierdzamy wciskając **Apply**.

3. Przygotowanie plików do analizy.

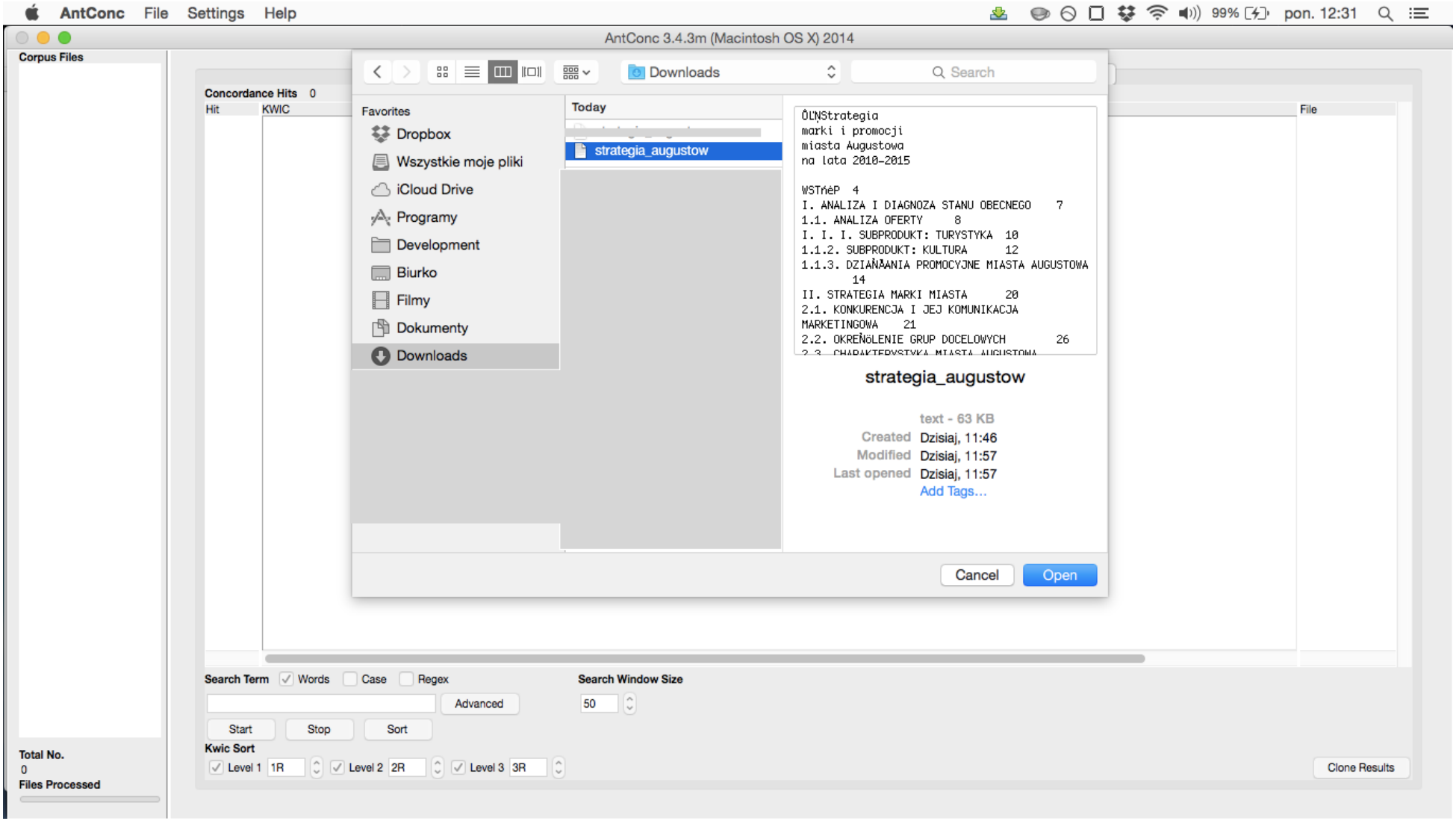
Program **Antoconc** pozwala na analizy korpusów tekstowych zapisanych jako proste pliki tekstowe. Pakiet **Microsoft Office** w trakcie wykonywanych czynności miał problem z eksportowaniem plików do formatu tekstowego. Dlatego też polecam wykorzystać w tym celu program Writer znajdujący się pakiecie **Libreoffice**. Program jest darmowy i licencjonowany na jednej z wolnych licencji.

Dokumenty strategiczne, które są udostępniane w formatach do edycji- np.: **docx,** czy **odt** należy po otwarciu zapisać jako dokument tekstowy. Natomiast dokumenty, które są dystrybuowane jako pliki **pdf**, należy poddać obróbce programem typu **OCR**, który rozpozna zawartą treść i pozwoli dokument zapisać jako plik do edycji.

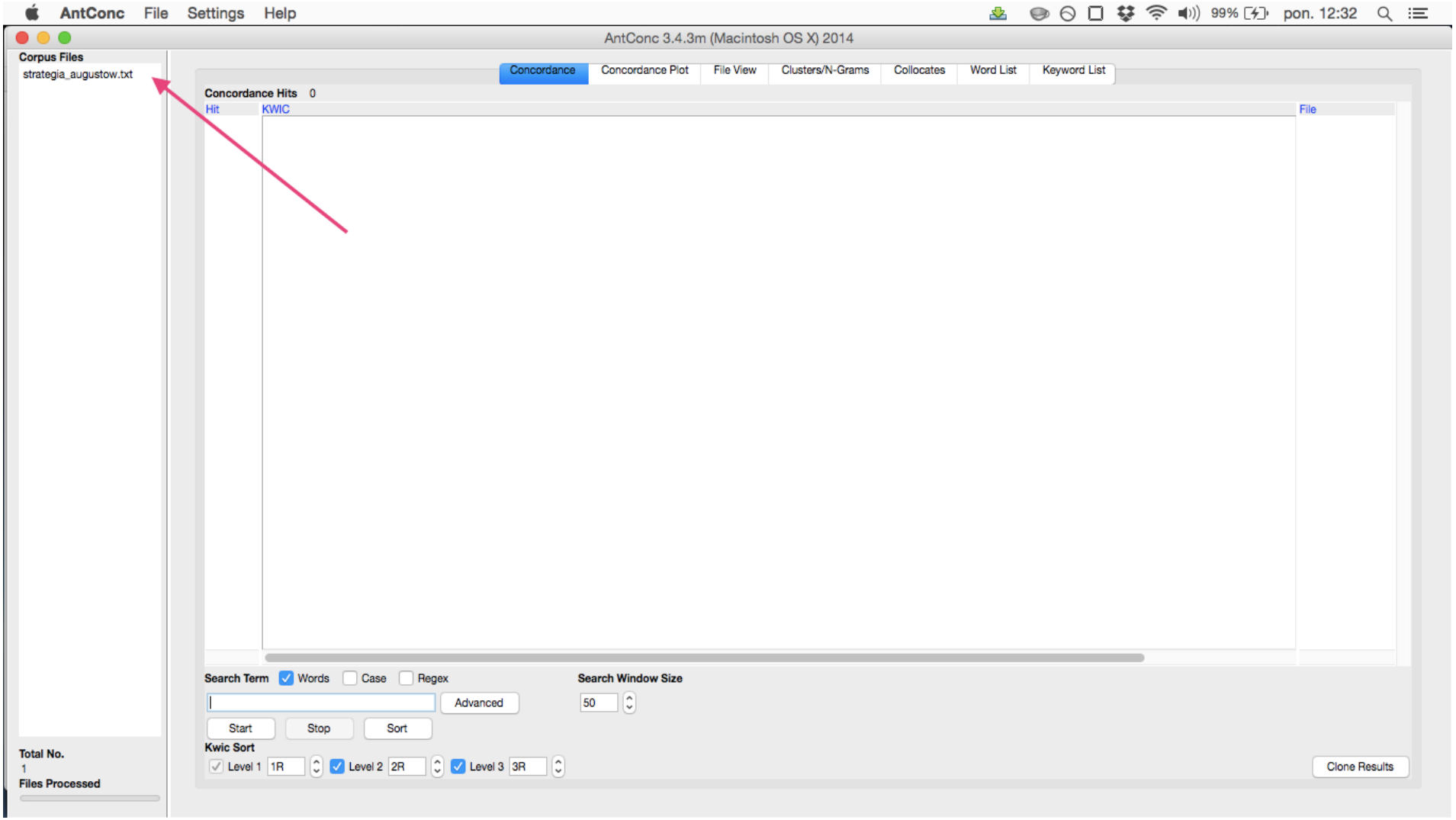
Tak przygotowany plik z korpusem tekstu możemy załadować do naszego programu. W tym celu wybieramy **File** i **Open File(s)**.



Pojawia się **manager plików**. Wybieramy stworzony przez nas plik z korpusem w formacie **txt**.



Poprawnie załadowany plik powinien pojawić się w **menu bocznym**.

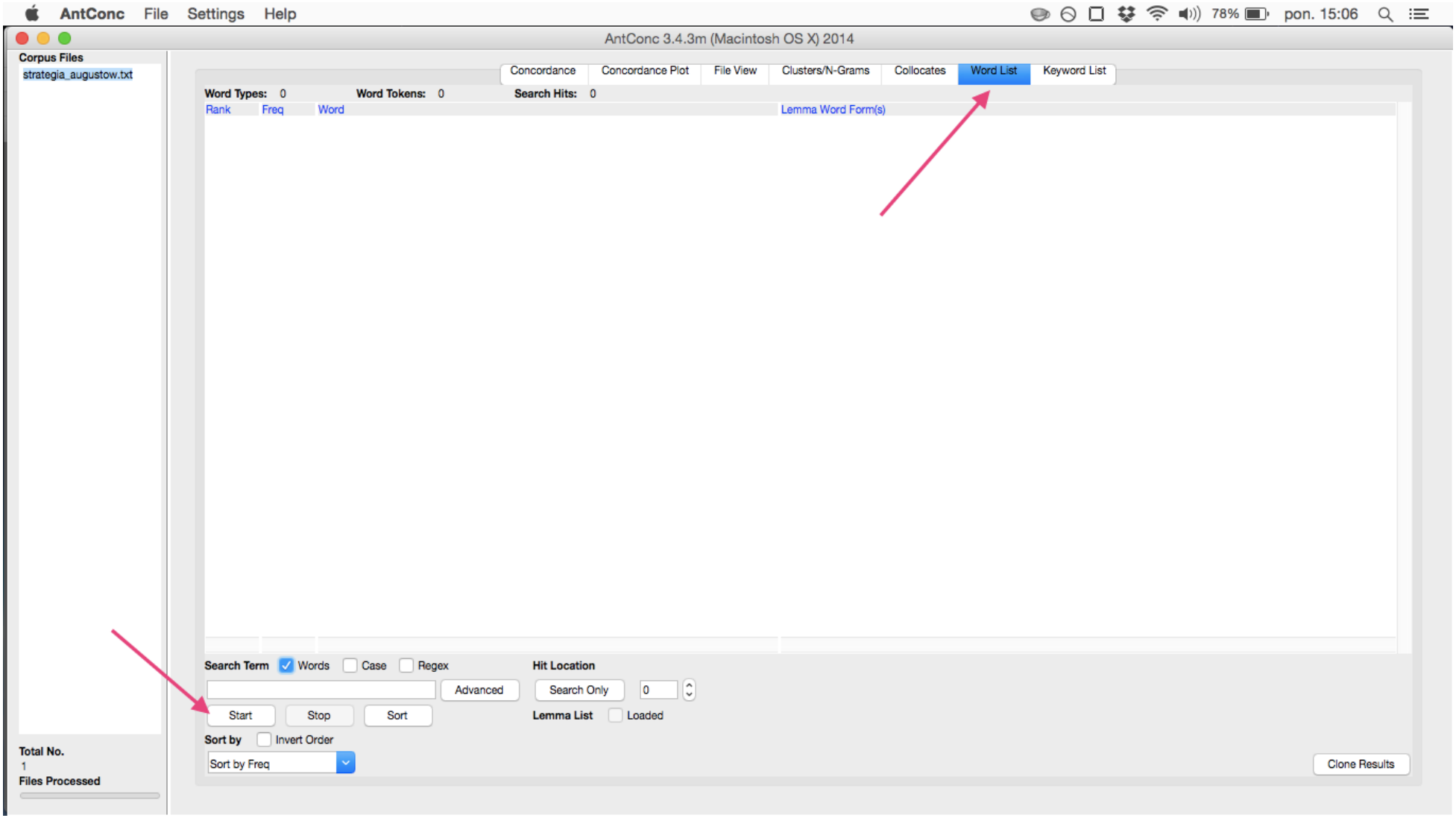


4. Analiza frekwencyjna.

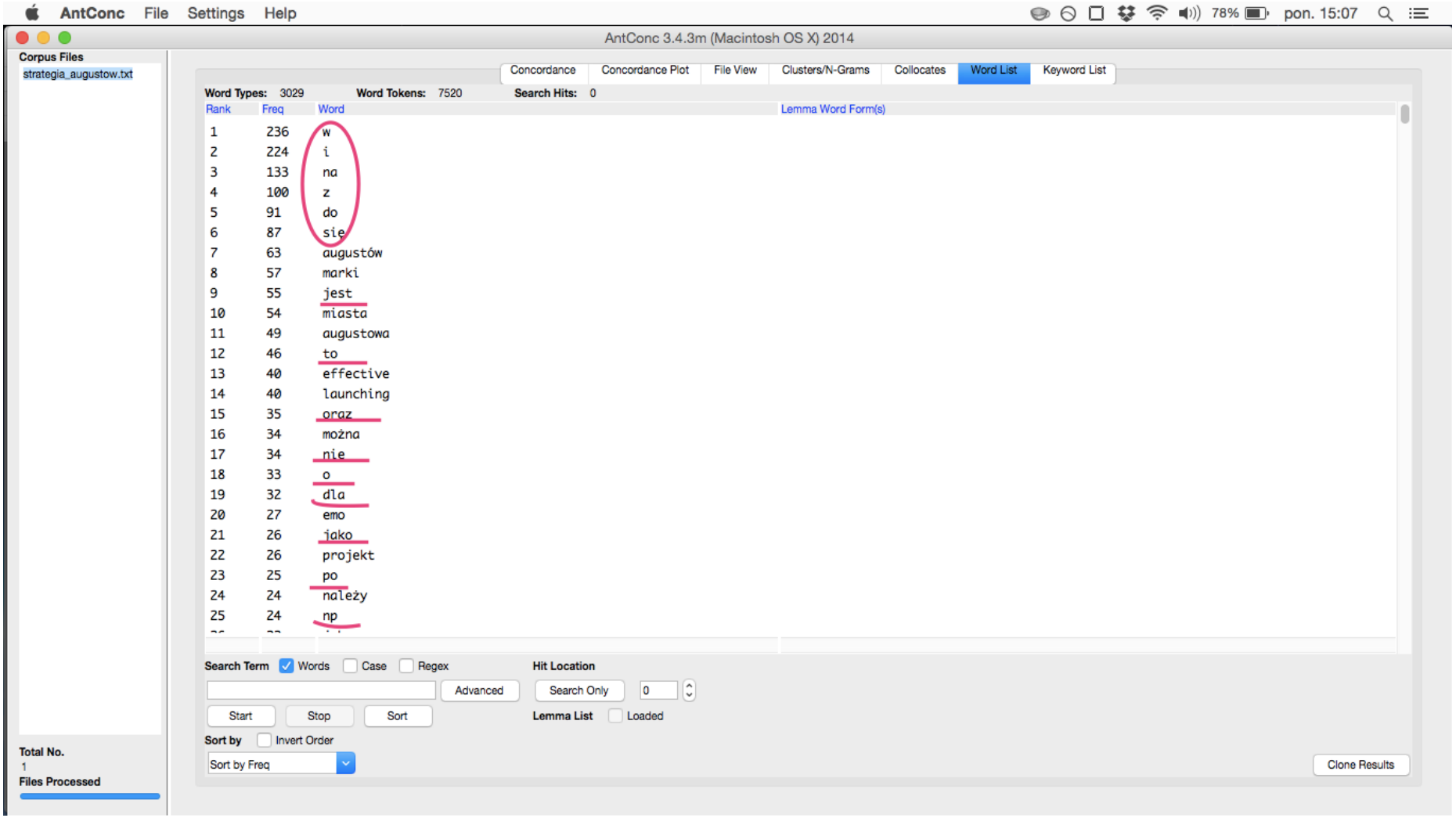
Po wykonaniu powyższych czynności można już przystąpić do pierwszych analiz.

Analiza frekwencyjna, wraz z wyszukiwaniem powtarzalnych połączeń wyrazowych (kolokacji) są podstawowymi technikami badawczymi na gruncie NLP (*Natural Language Processing*) dziedziny korzystającej z językoznawstwa i informatyki.

Pierwszym krokiem jest przejście do zakładki **Word List**. I następnie wciśnięcie przycisku **Start**.

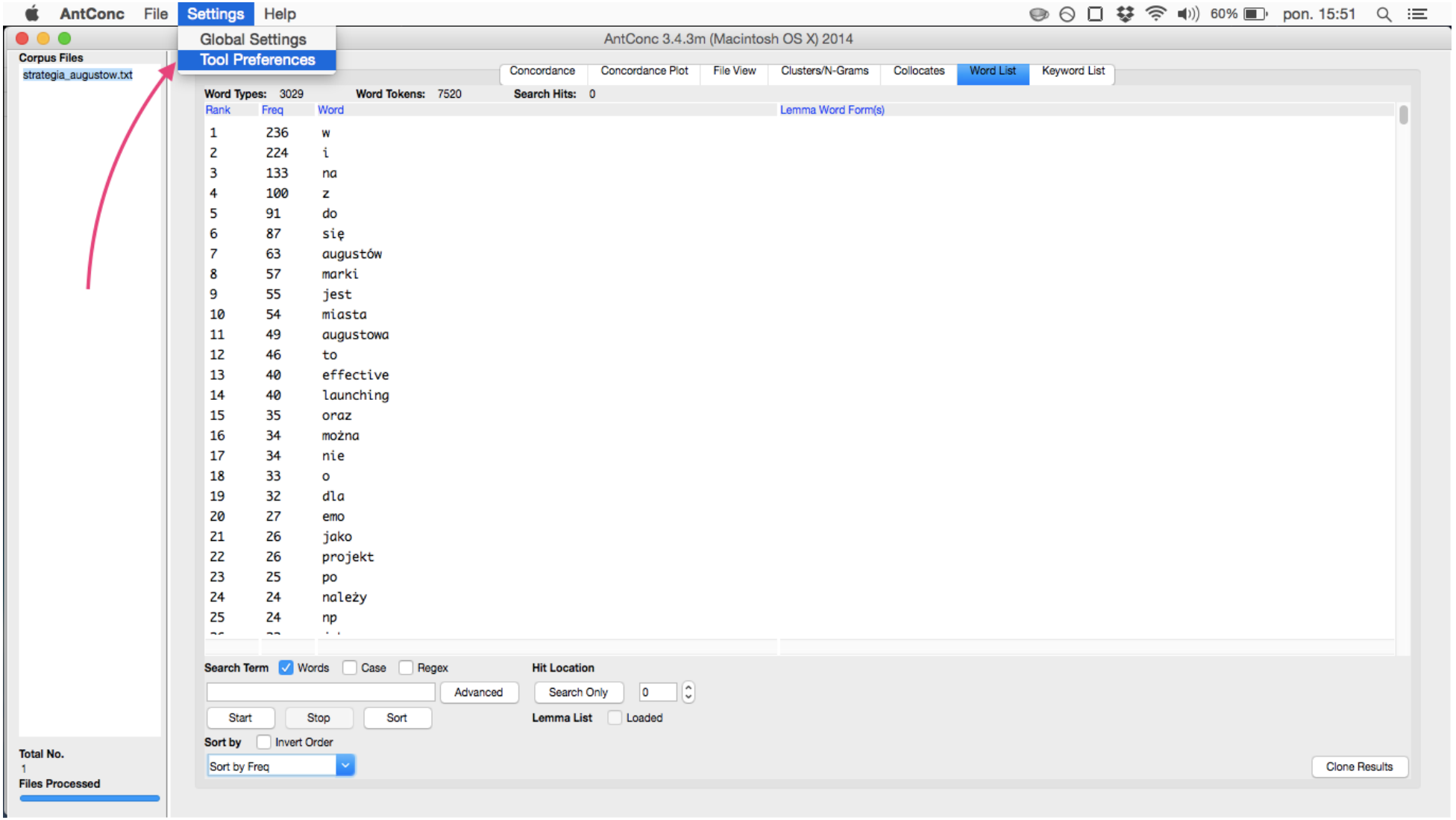


W efekcie otrzymujemy surową listę frekwencyjną słów z korpusu tekstowego. W tekście eliminujemy zaimki, spójniki i przyimki. Z punktu widzenia naszych analiz nie noszą one interesujących treści.

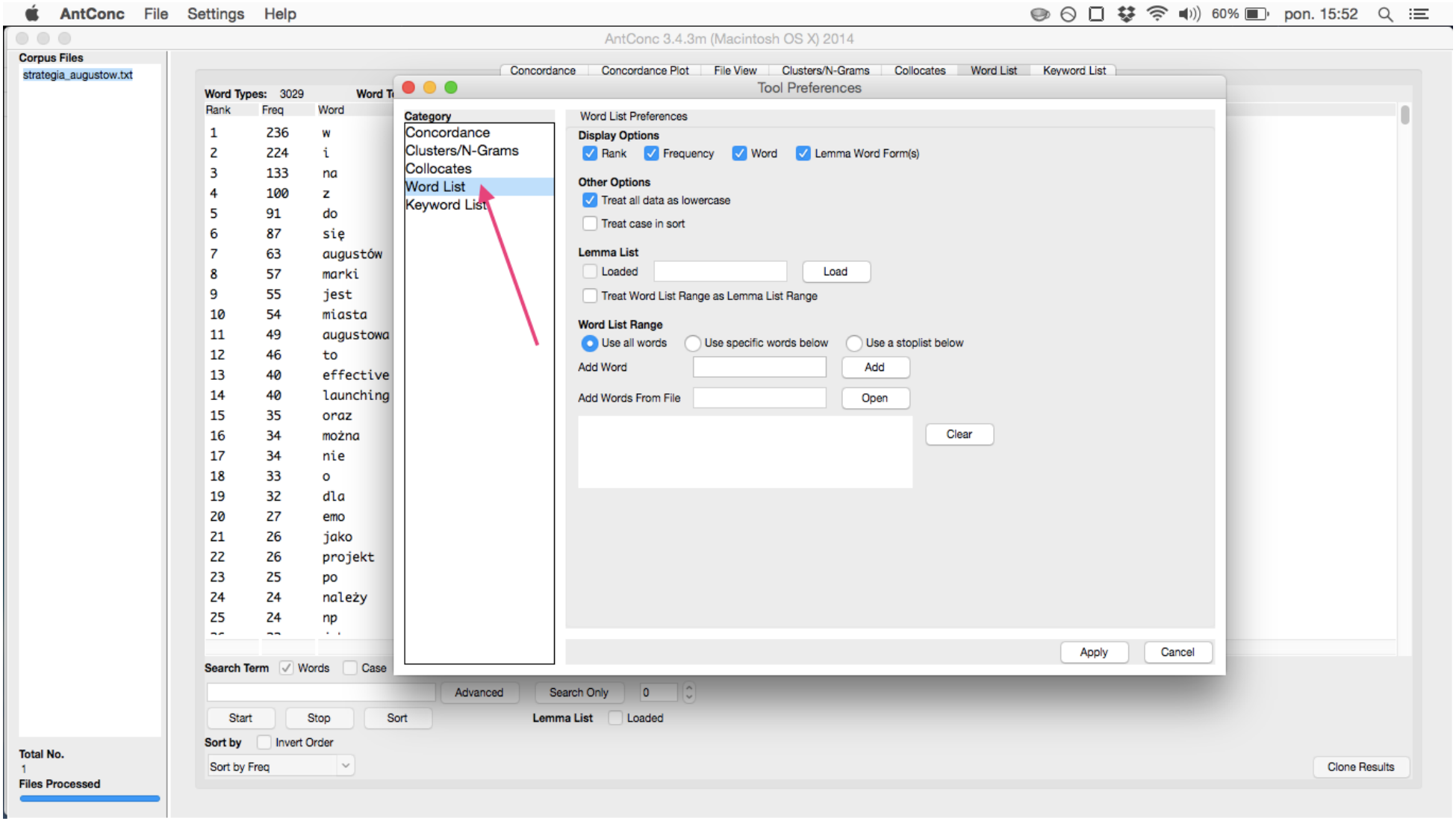


Aby wyczyścić tekst z tego typu niepożądanych słów można skonstruować **Stoplistę**. Jest to zbiór słów, które mają być wyłączone z procesu przetwarzania i nie będą się pojawiać w dalszych analizach. Przykładowa stoplista znajduje się w zasobach wikipedii (http://pl.wikipedia.org/wiki/Wikipedia:Stopwords). Proszę kliknąć w plik z naszą stoplistą .
Struktura pliku jest bardzo prosta, słowa są oddzielone przecinkami. Można więc ten plik dowolnie rozbudowywać, bądź też usuwać słowa, jeśli uznamy, że jest on zbyt rozbudowany.

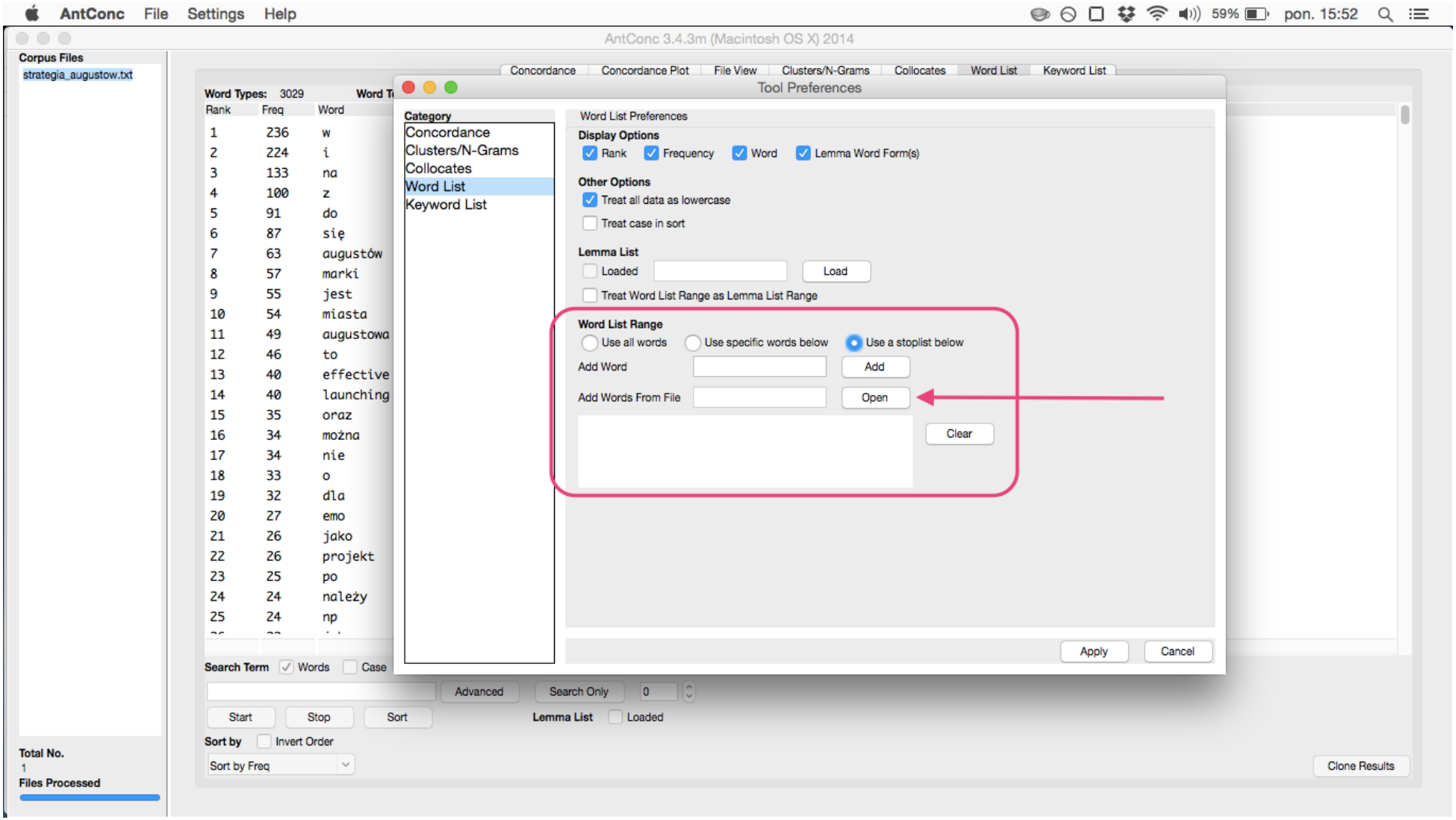
Aby załadować **stoplistę** należy w menu głównym wybrać **Settings** i następnie **Tool Preferences.**



W oknie, które się pojawia wybieramy **Word List.**

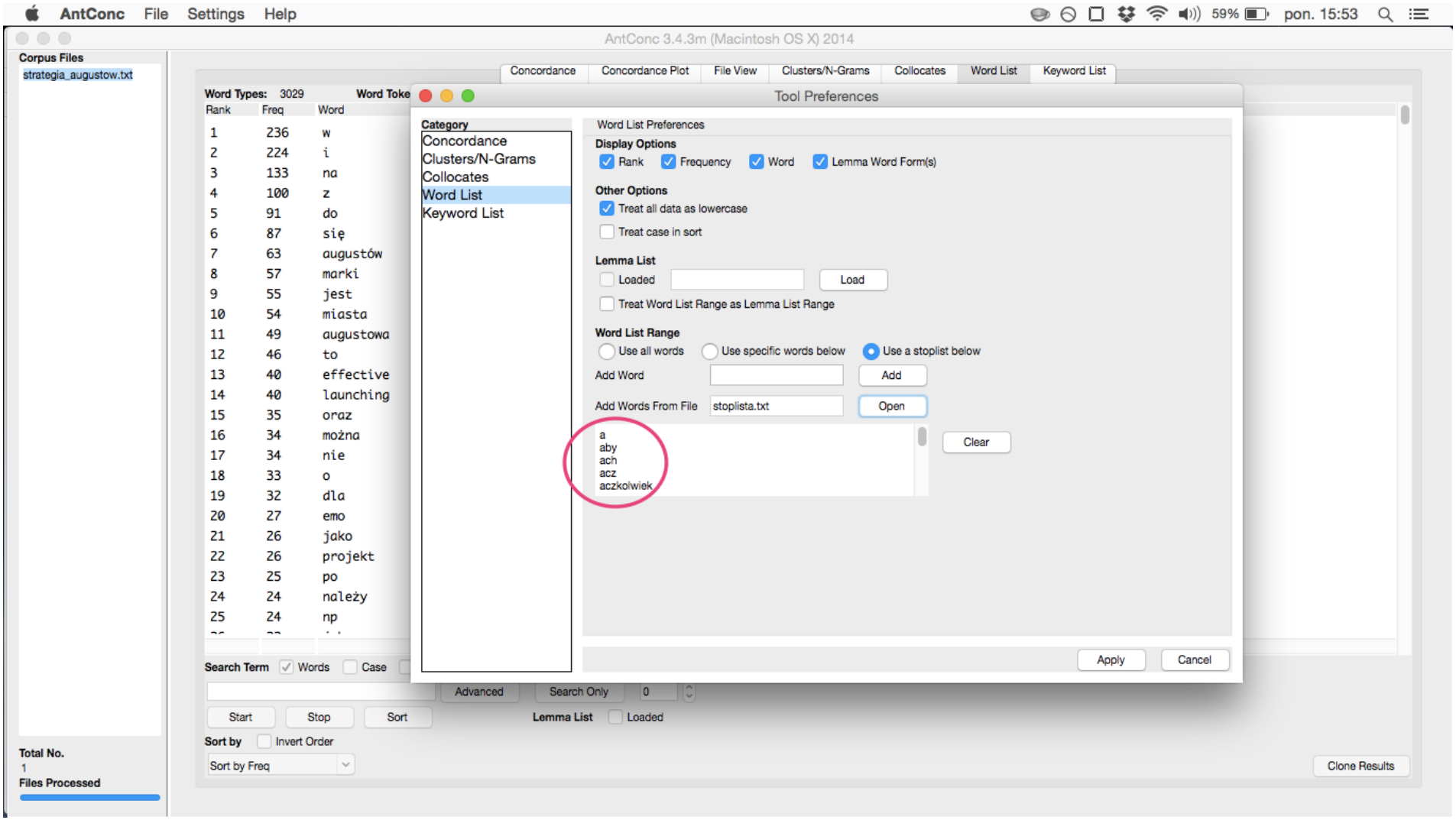


Przechodzimy do fragmentu **Word List Range**. Ustawiamy wartość **Use a stoplist below** i w opcji **Add Words From File** naciskamy **Open**. Wybieramy nasz plik **stoplist.txt**.

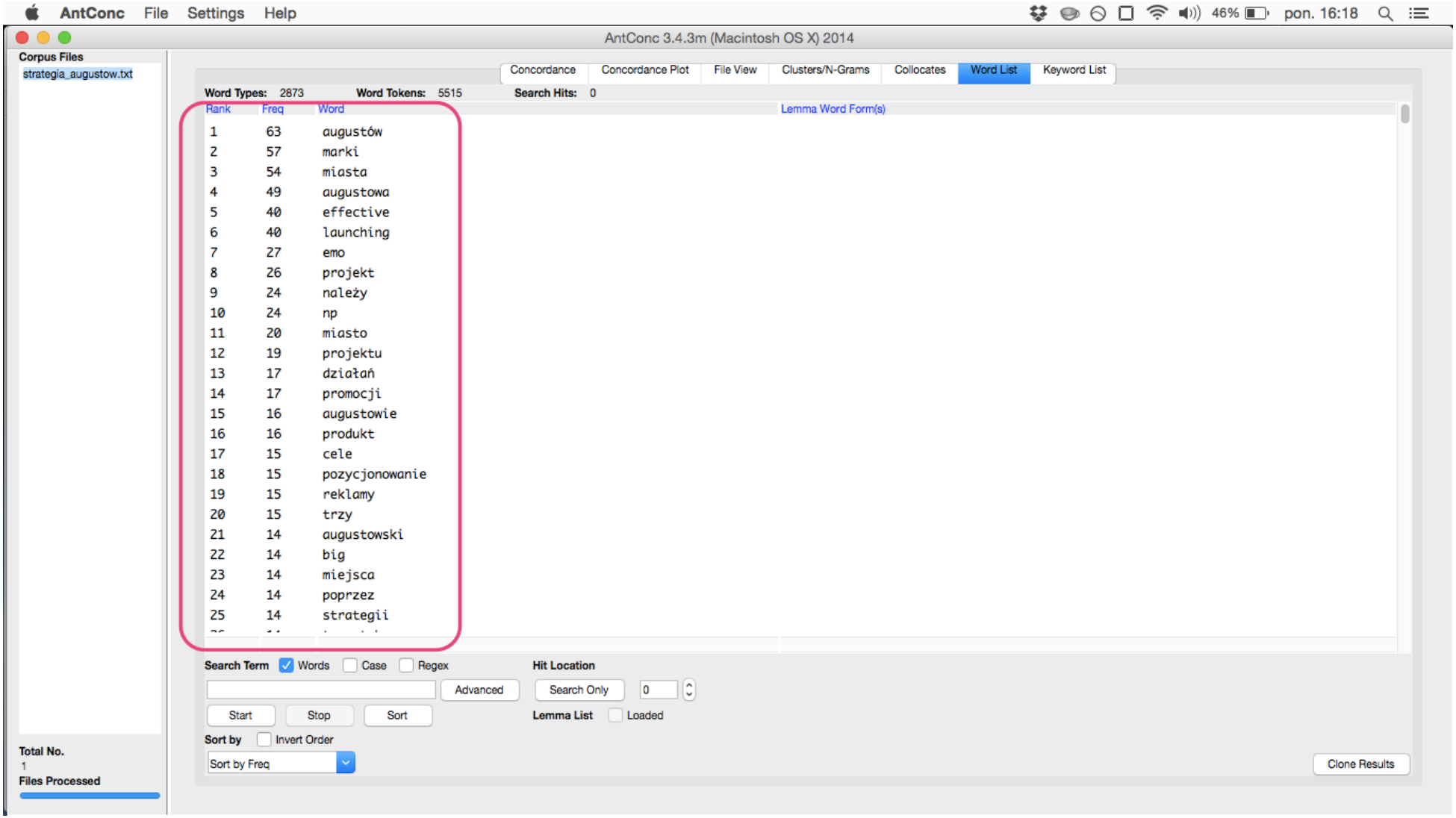


Jeśli czynność zostanie zakończona sukcesem słowa ze stoplisty powinny się załadować do programu.

Wybór potwierdzamy przyciskiem **Apply.**



Należy następnie jeszcze raz przeładować listę, dopiero wówczas pojawią się frekwencje słów, bez słów ze stoplisty.



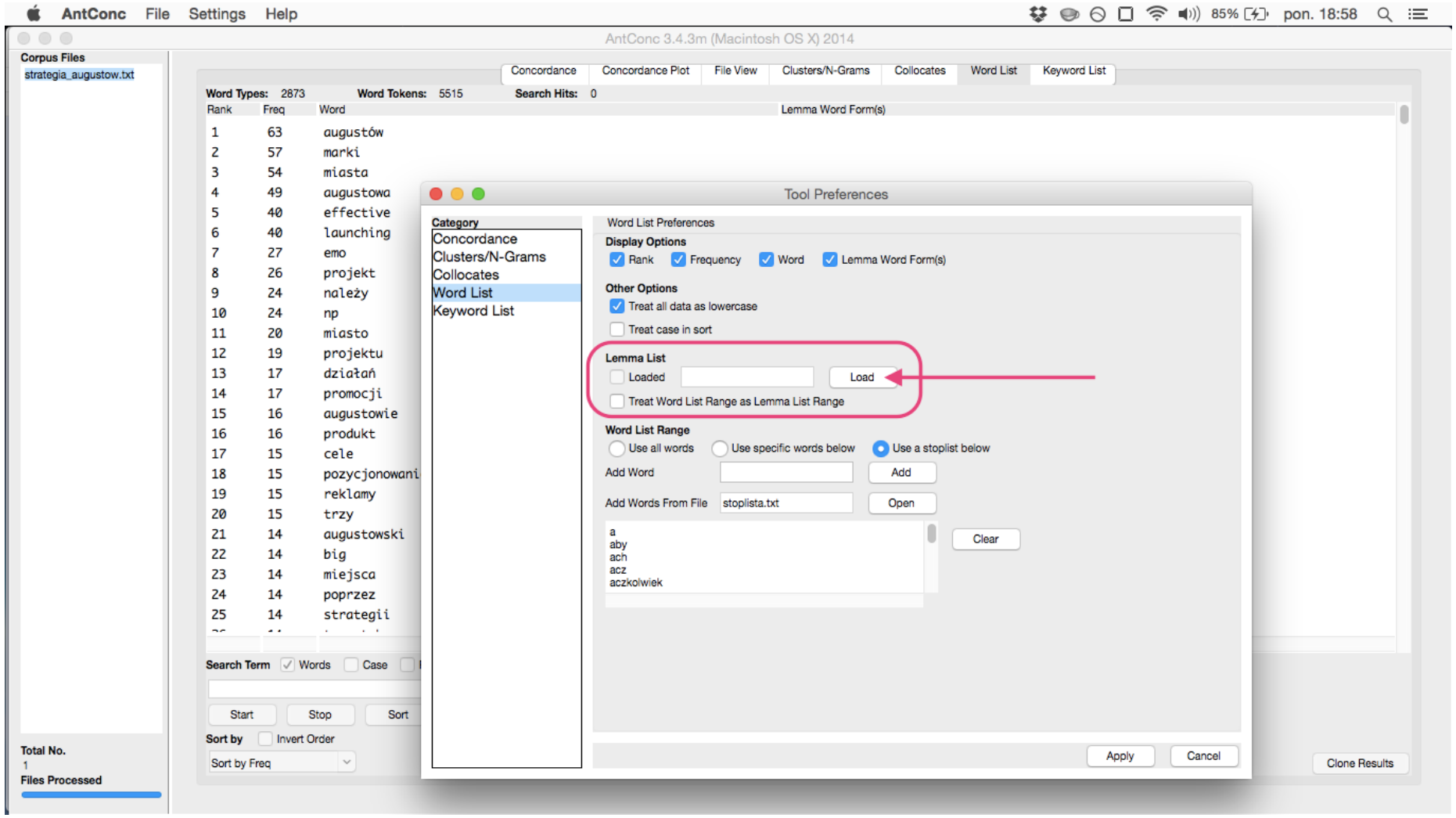
Praca ze słownikiem języka polskiego

Tak wyodrębnione frekwencje nie uwzględniają jednak różnorodnych końcówek fleksyjnych. Jak widać na wcześniejszym obrazie - 63 razy pojawia się Augustów, 49 - Augustowa, 16 - Augostowie, 14 - augustowski. W celu właściwego określenia frekwencji korpusu tekstowego należy zliczyć więc formy podstawowe słów, w językoznawstwie zwane - *lemma*. Osobno zostaną zliczone rzeczowniki, czasowniki i przymiotniki. Język polski jest językiem fleksyjnym, z dużą liczbą wyjątków, co sprawia, że nie można zastosować procesu *stemmingu*, czyli programów tworzących formy fleksyjne w oparciu o ustalone reguły.

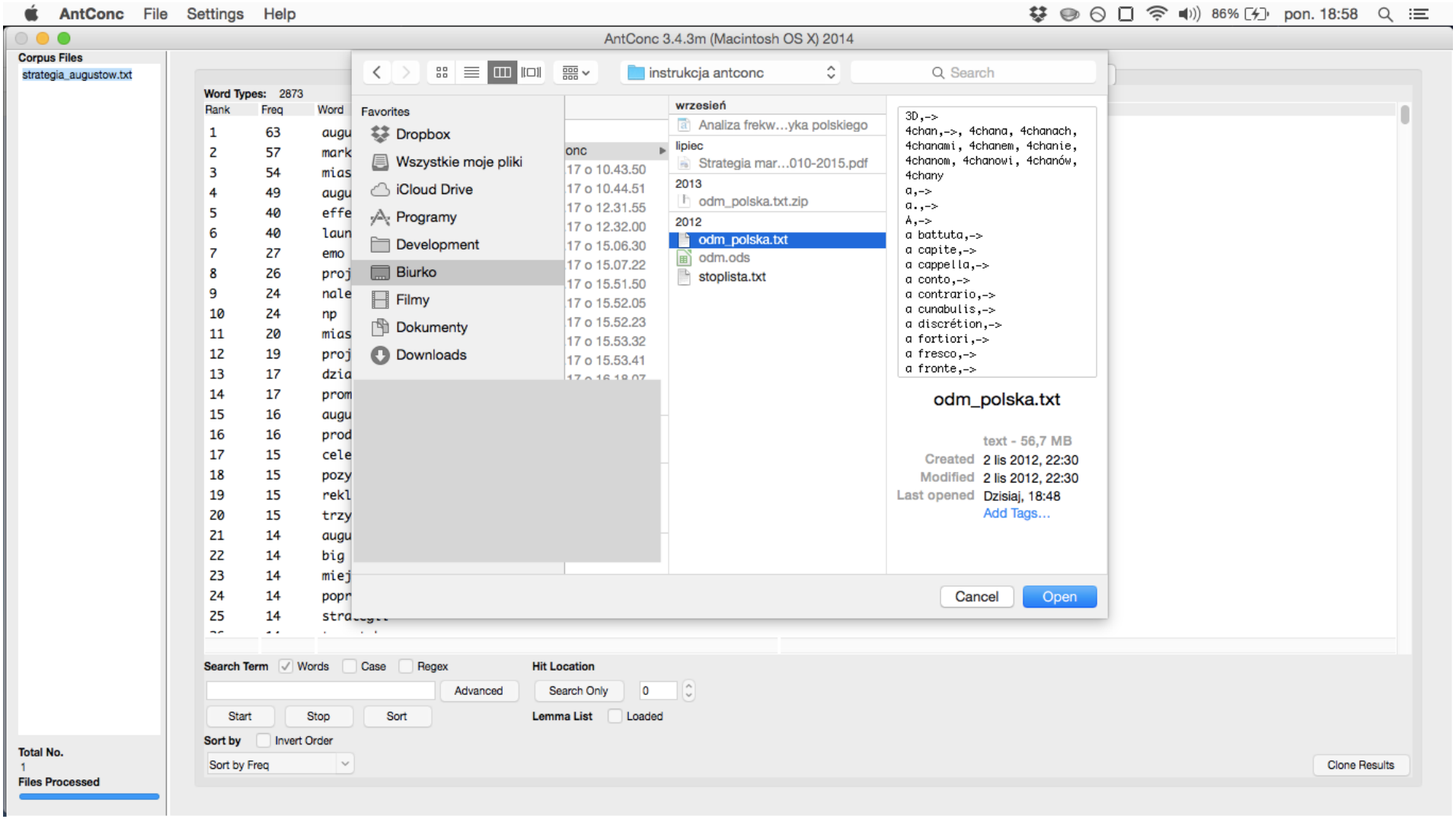
W celu obejścia tego problemu można posłużyć się przygotowanym słownikiem języka polskiego. Słownik ten został przygotowany w oparciu o zasoby **aspell**. Słownik znajduje się w **zakładce teksty ożk**, wraz z plikiem stoplista.txt. Składa się on dotąd ze 194 tys. form podstawowych, i ponad 3,6 mln słów w różnych formach fleksyjnych.

Uwaga: ze względu na rozmiar pliku ze słownikiem- ponad 55 Mb, do pracy rekomendowany jest komputer posiadający co najmniej 4 Gb pamięci operacyjnej (RAM).

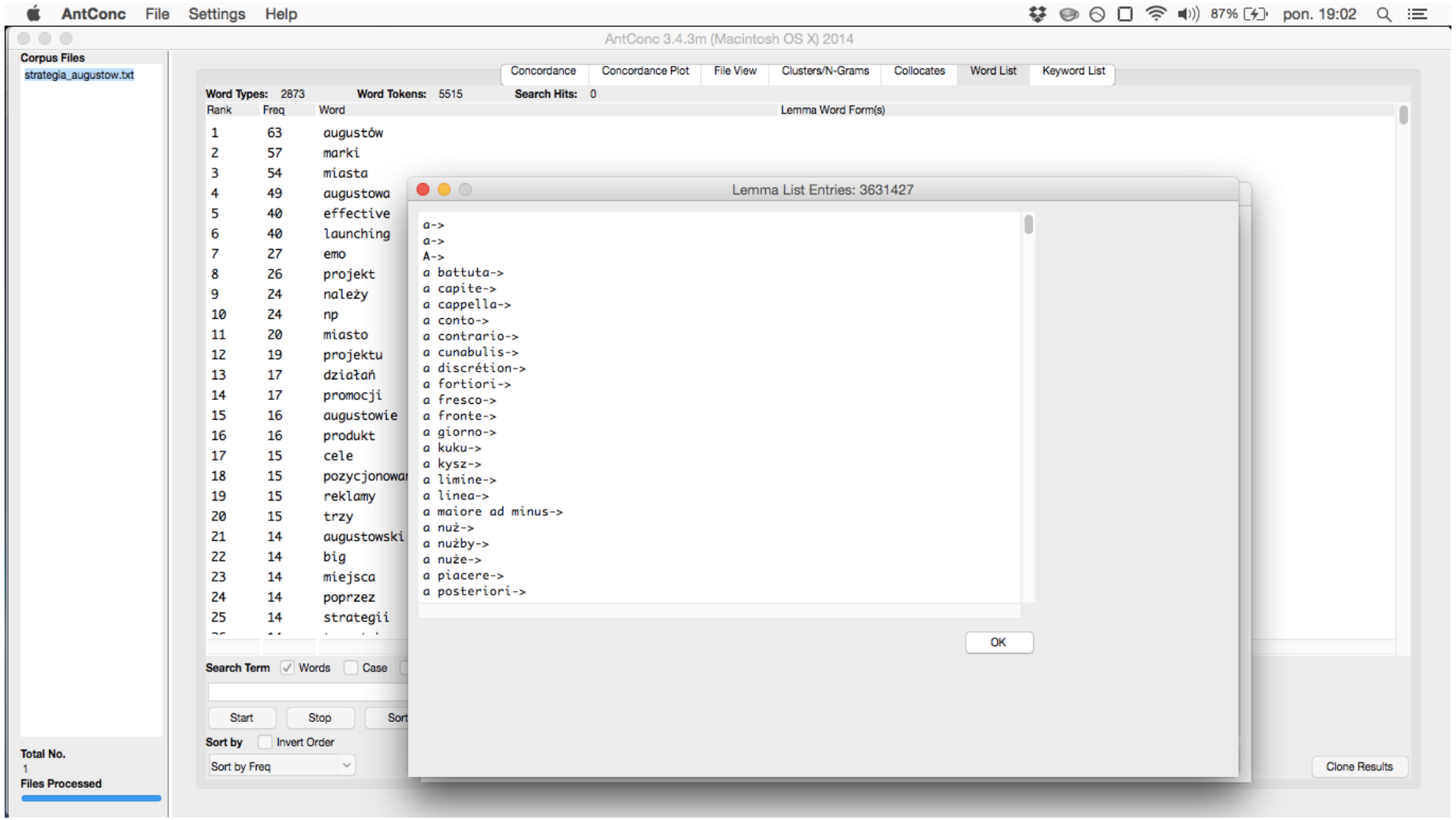
Aby załadować słownik języka polskiego do programu **Antconc** należy przejść do menu głównego i wybrać **Settings** i **Tool Preferences**, (por. strona 13) i przejść do opcji **Word List** (por. strona 14). W obszarze **Lemma List** wciskamy **Load** i



wczytujemy nasz słownik języka polskiego. Proces ten jest czasochłonny, i może trwać kilka minut, w zależności od szybkości komputera. W tym czasie program może nie odpowiadać.

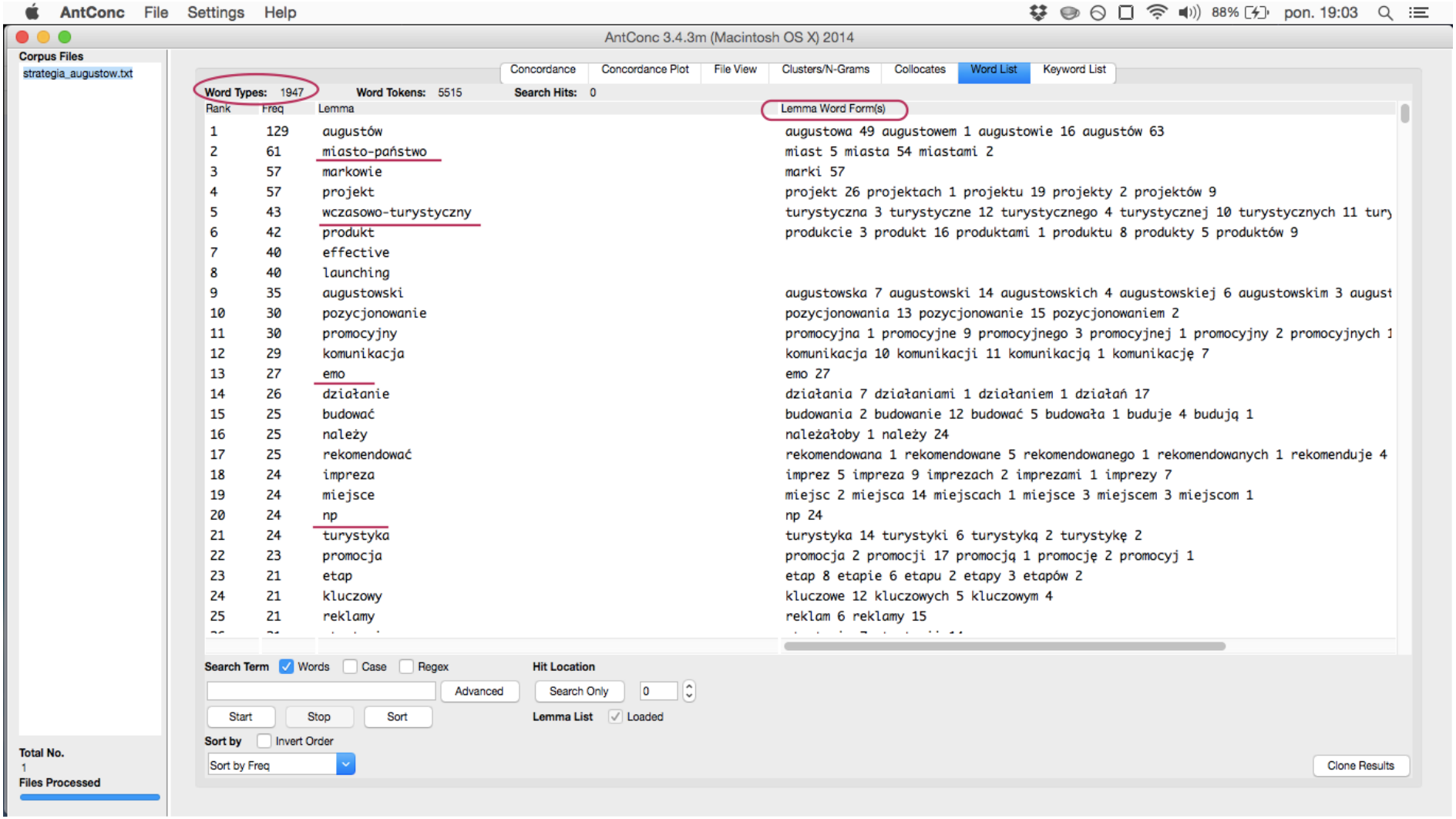


Jeśli operacja przebiegnie pomyślnie ukaże się ekran, z formami fleksyjnymi. Teraz wystarczy nacisnąć **OK,** a następnie



nacisnąć **Apply**.

Pozostaje jeszcze ponownie wygenerować listę słów - przyciskiem **Start**. Efekt jest następujący.



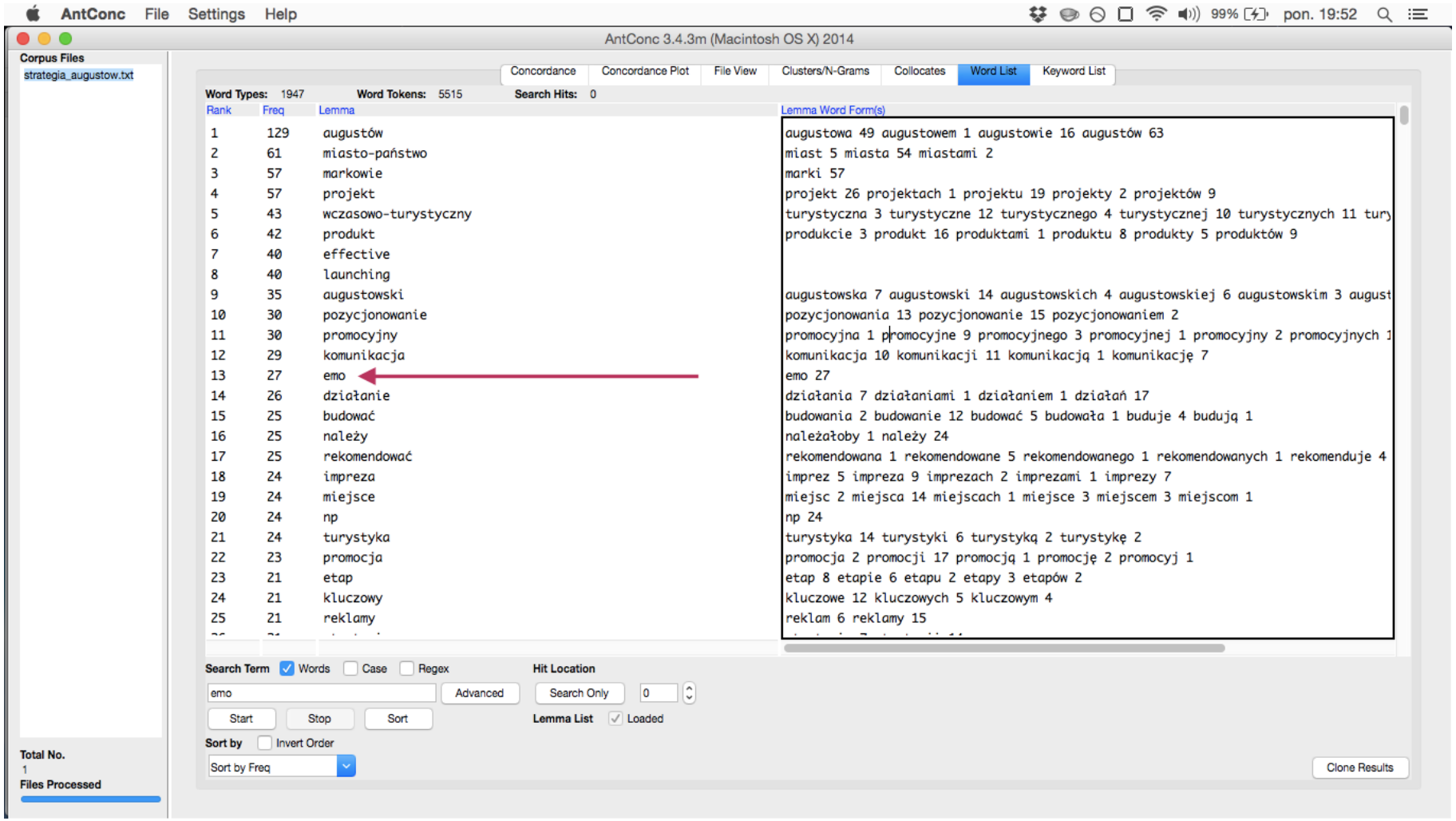
Dzięki zastosowaniu słownika języka polskiego oraz stopliście udało się zredukować listę słów o 1000 pozycji. Wynosi ona teraz 1947 pozycji (na początku było to 3029 słów). Jednak całego procesu nie można zautomatyzować. Pojawiają się błędy, które wynikają z samej natury automatyzacji procesów analizy językowej.

Częstym typem błędu jest pojawienie się wyrazów połączonych myślnikiem, mimo tego, że w samym tekście takie zwroty nie funkcjonowały. W analizowanym korpusie było to *miasto-państwo* i *wczasowo-turystyczny*. Natomiast dość łatwo można dojść do źródeł tych słów - podgląd form, które zsumowały się na liczebność formy podstawowej można zobaczyć w głównym oknie w kolumnie **Lemma Word Form(s)**. I można przekonać się, że *miasto-państwo* to faktycznie rożne formy rzeczownika *miasto*, *a wczasowo-turystyczny* to różne formy przymiotnika -*turystyczny*.

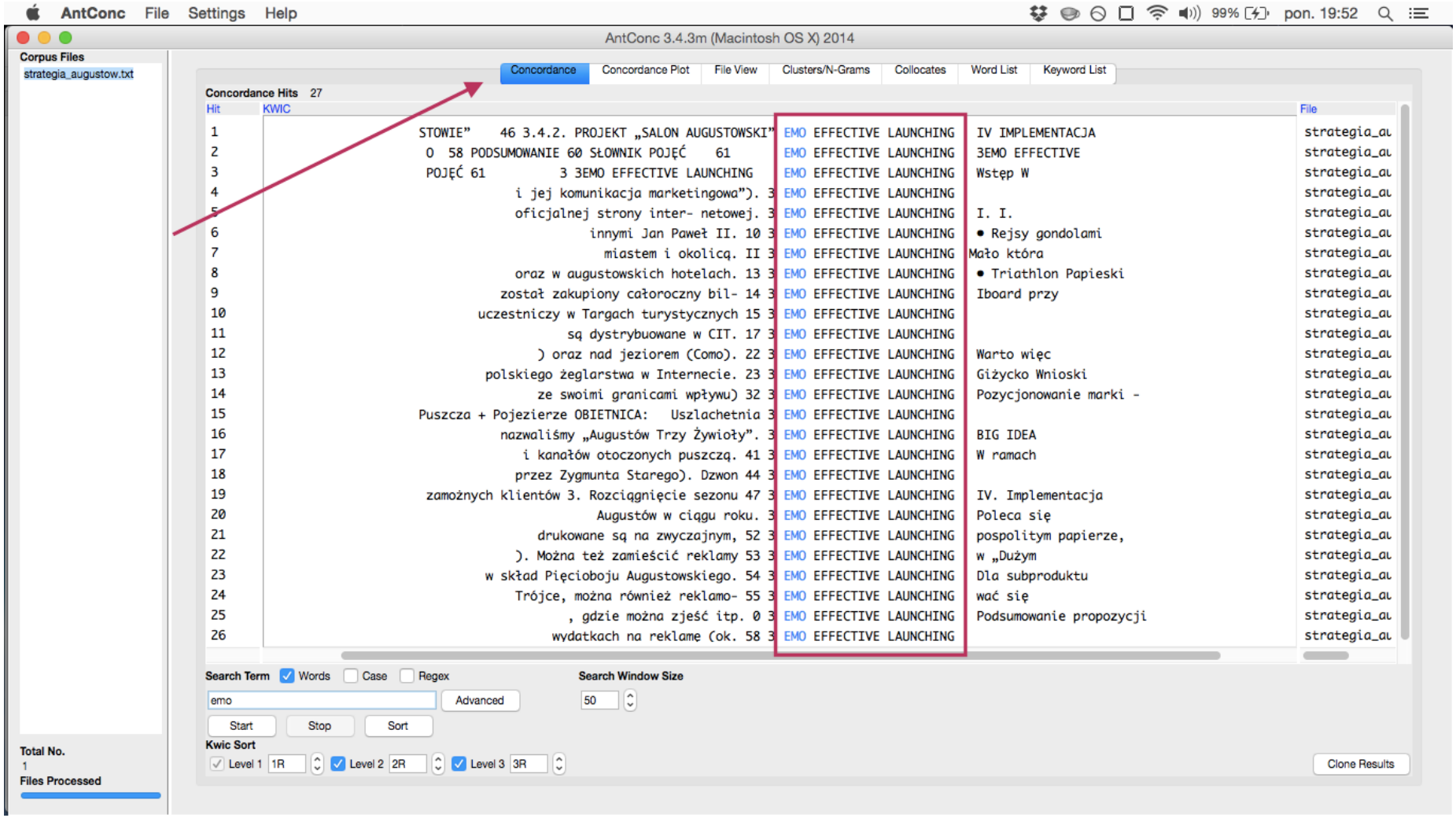
Natura automatyzacji procesu analizy językowej sprawia, że pojawiają się nieprawidłowo dobrane ze względu na funkcję słowa formy podstawowe. Na 3 miejscu pojawiło się słowo *markowie* - jako lemma słowa *marki* (z kontekstu tekstu wynika, że jest to wyraz *marka* - jako znak rozpoznawczy produktu w liczbie mnogiej). Jest to aspekt, który wymaga od badacza dużej czujności w trakcie wyciągania wniosków o naturze tekstu.

Wreszcie mogą pojawiać się niezrozumiałe w pierwszym momencie ciągi znaków. Program Antconc pozwala na szybki podgląd słowa, i określenie jego kontekstu.

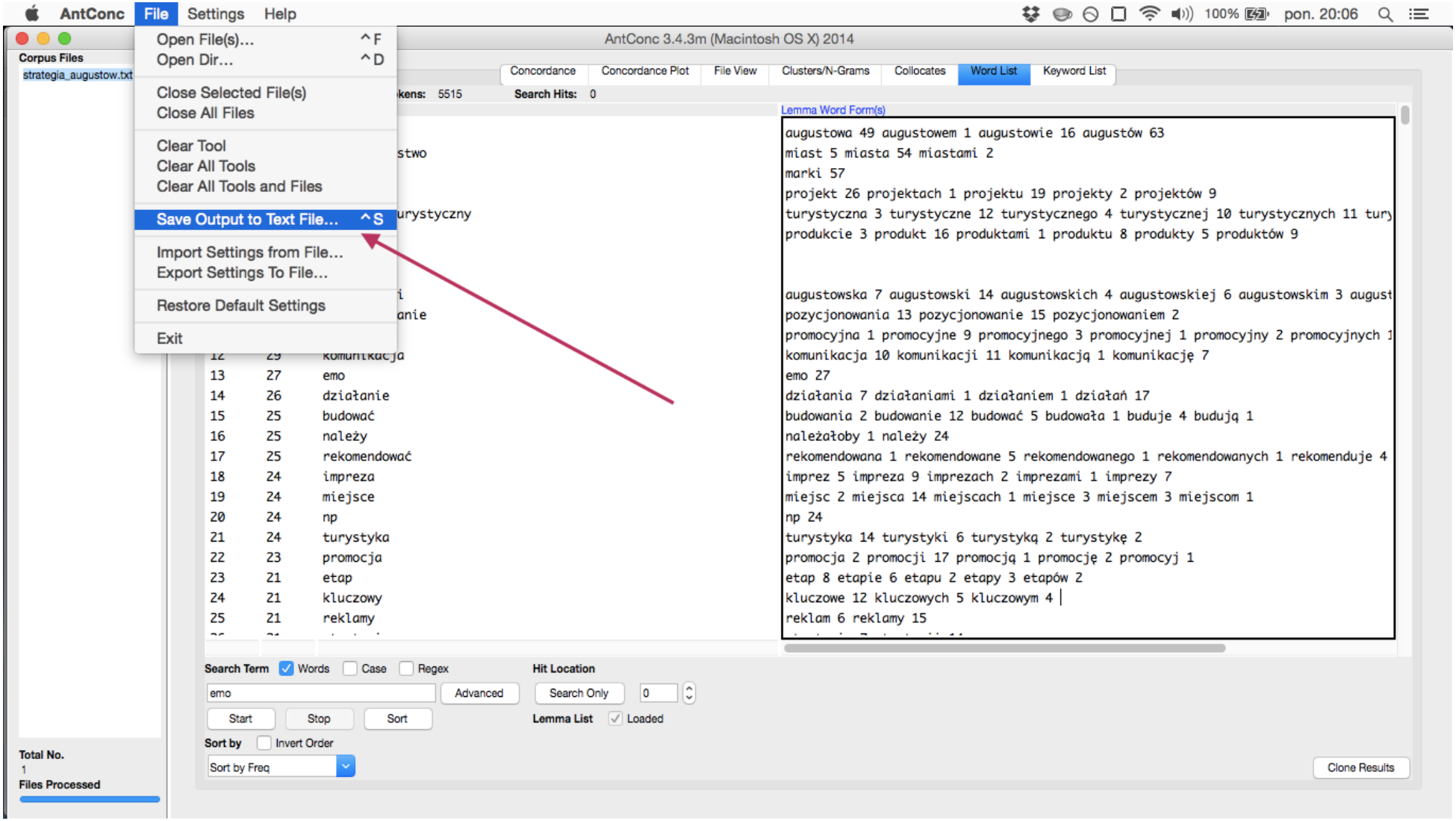
Słowo, które chcemy zbadać można kliknąć. Wystarczy w tym celu wybrać słowo w kolumnie Lemma. Np. *emo*.



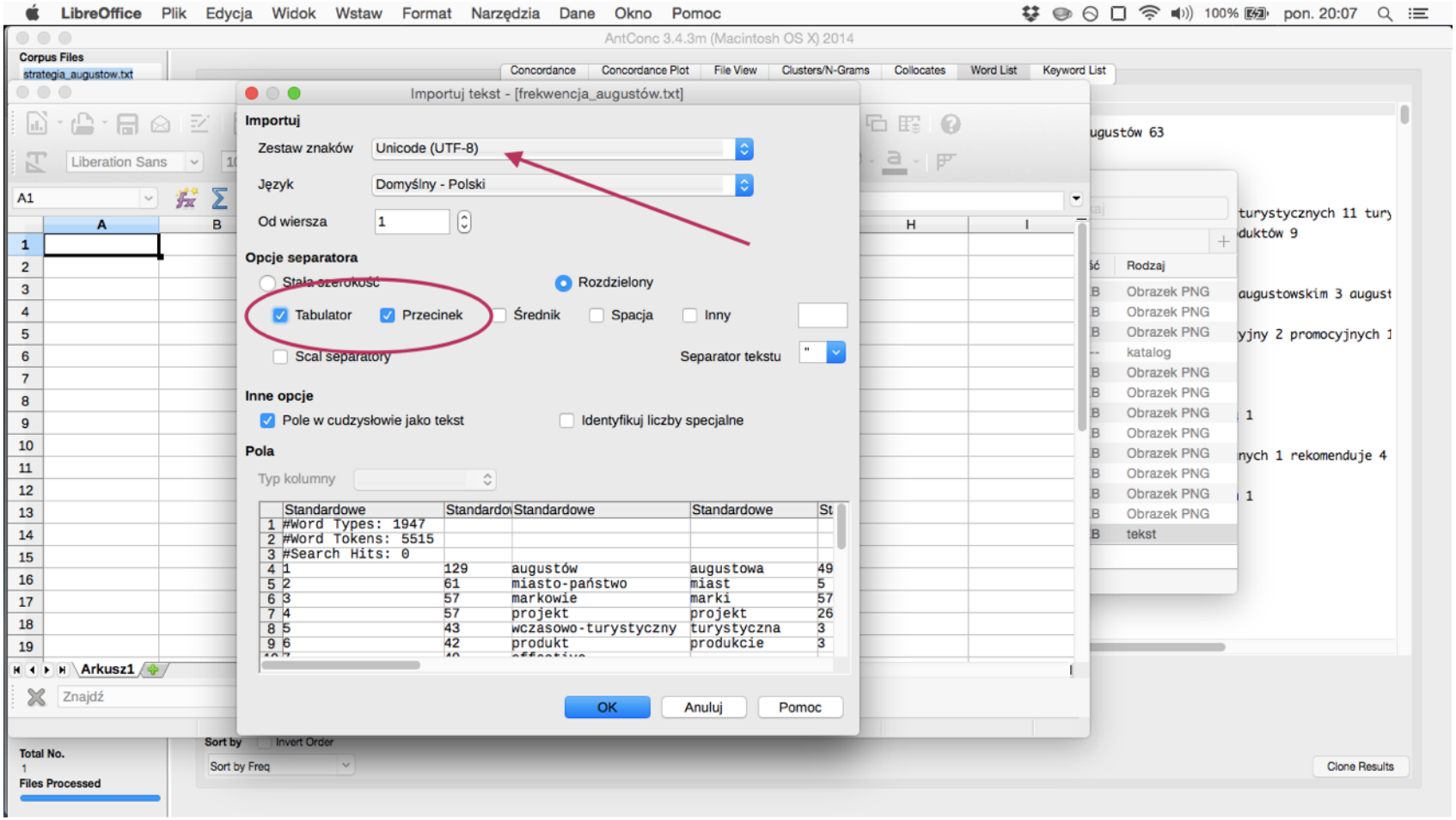
Program przeniesie nas do zakładki **Concordance**, w której można podejrzeć kontekst wyrazu. Nasz wyraz jest fragmentem nazwy *EMO EFFECTIVE LAUNCHING*. Tak nazywa się firma projektująca strategię, która na każdej stronie dokumentu zaznaczyła swoją obecność. Więcej o tej opcji znajduje się w rozdziale 6.



Aby móc wyeksportować plik z frekwencjami wyrazów do arkusza kalkulacyjnego należy wrócić do zakładki **Word List**, wybrać z menu głównego **File** i **Save Output to Text File.** Następnie zapisujemy plik z wynikami.



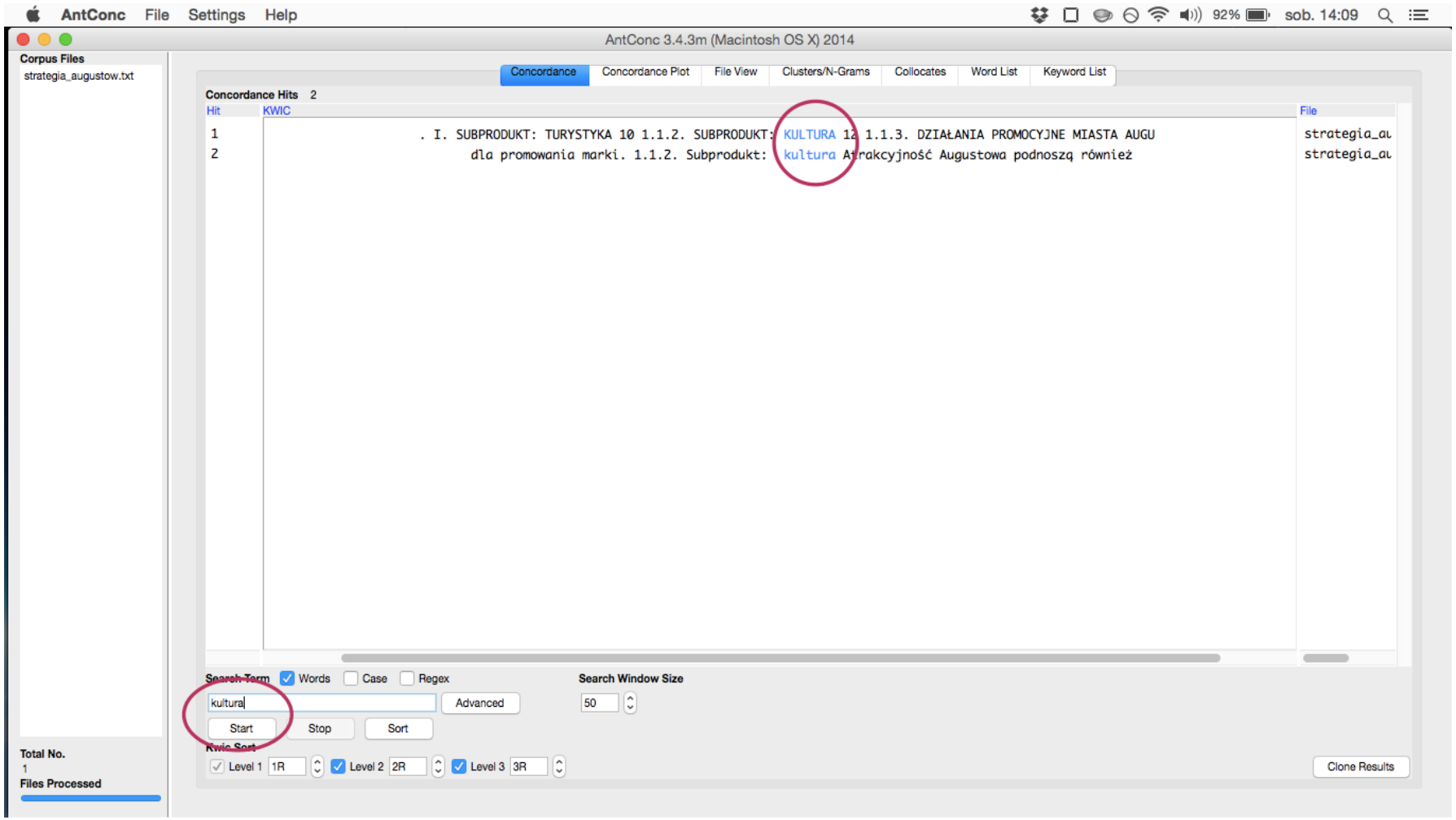
Plik wynikowy jest plikiem tekstowym, który bardzo łatwo można otworzyć w arkuszu kalkulacyjnym. Do tego celu polecam użycie **programu Calc z pakietu Libreoffce**. W programie Calc otwieramy nasz plik i pojawia się opcja jego importu. Należy ustawić opcję kodowania znaków na **UTF-8**, oraz w opcjach separatora zaznaczyć - **Tabulator** i **Przecinek**. Następnie można już dać **OK.**



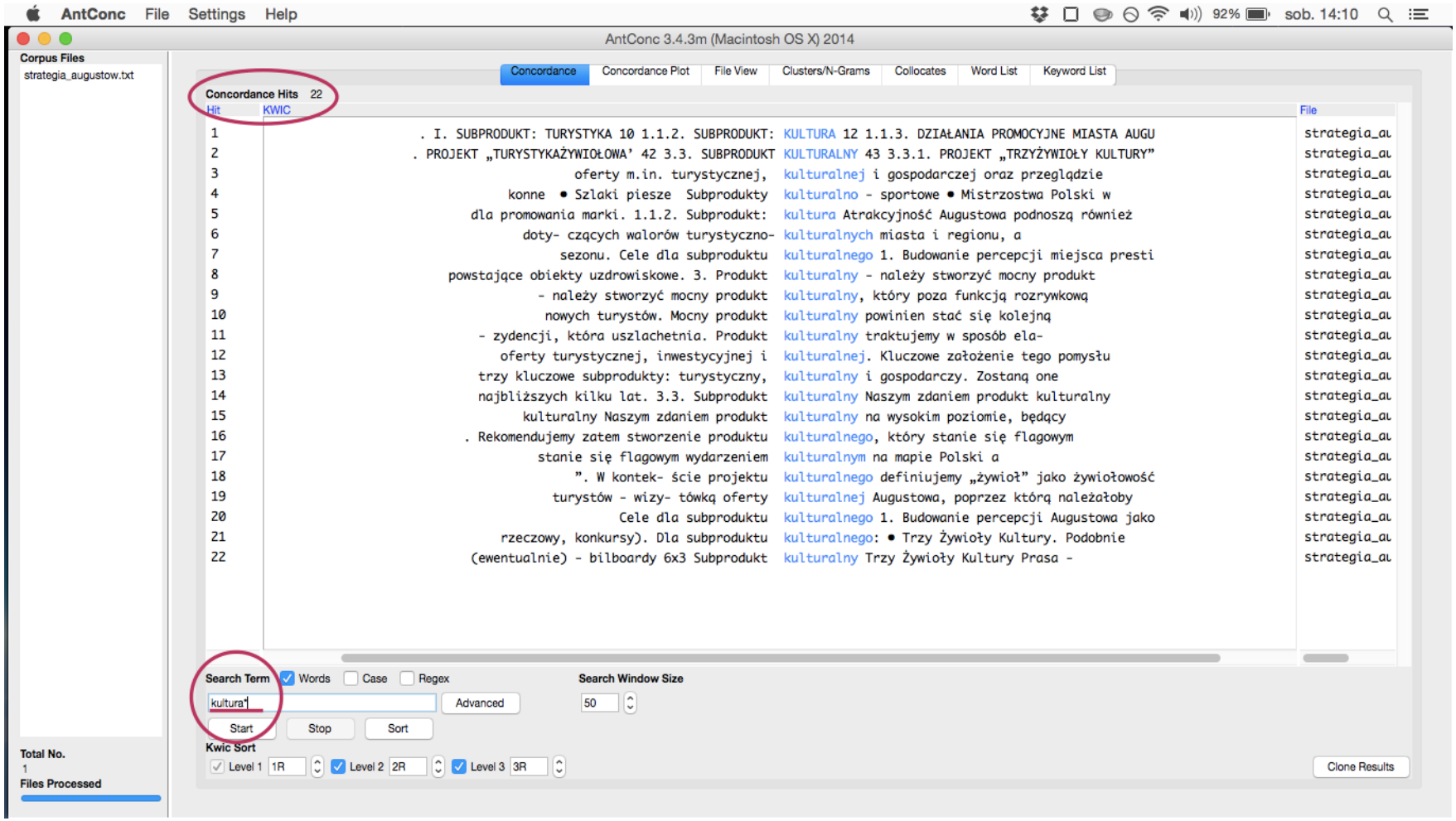
5. Słowo w kontekście.

Concordance jest domyślną zakładka programu **Antconc**. Otwiera się ona natychmiast po uruchomieniu. Dostępne tu funkcje mogą służyć jako narzędzia do weryfikacji automatycznej analizy frekwencyjnej, mogą pomóc w poprawianiu błędów. Badacz posiada pełną kontrolę nad procesem. Natomiast najciekawszą funkcją jest zaawansowany **moduł wyszukiwania pełnotekstowego**. Jego funkcjonalność zostanie zademonstrowana na przykładzie słowa *kultura*.

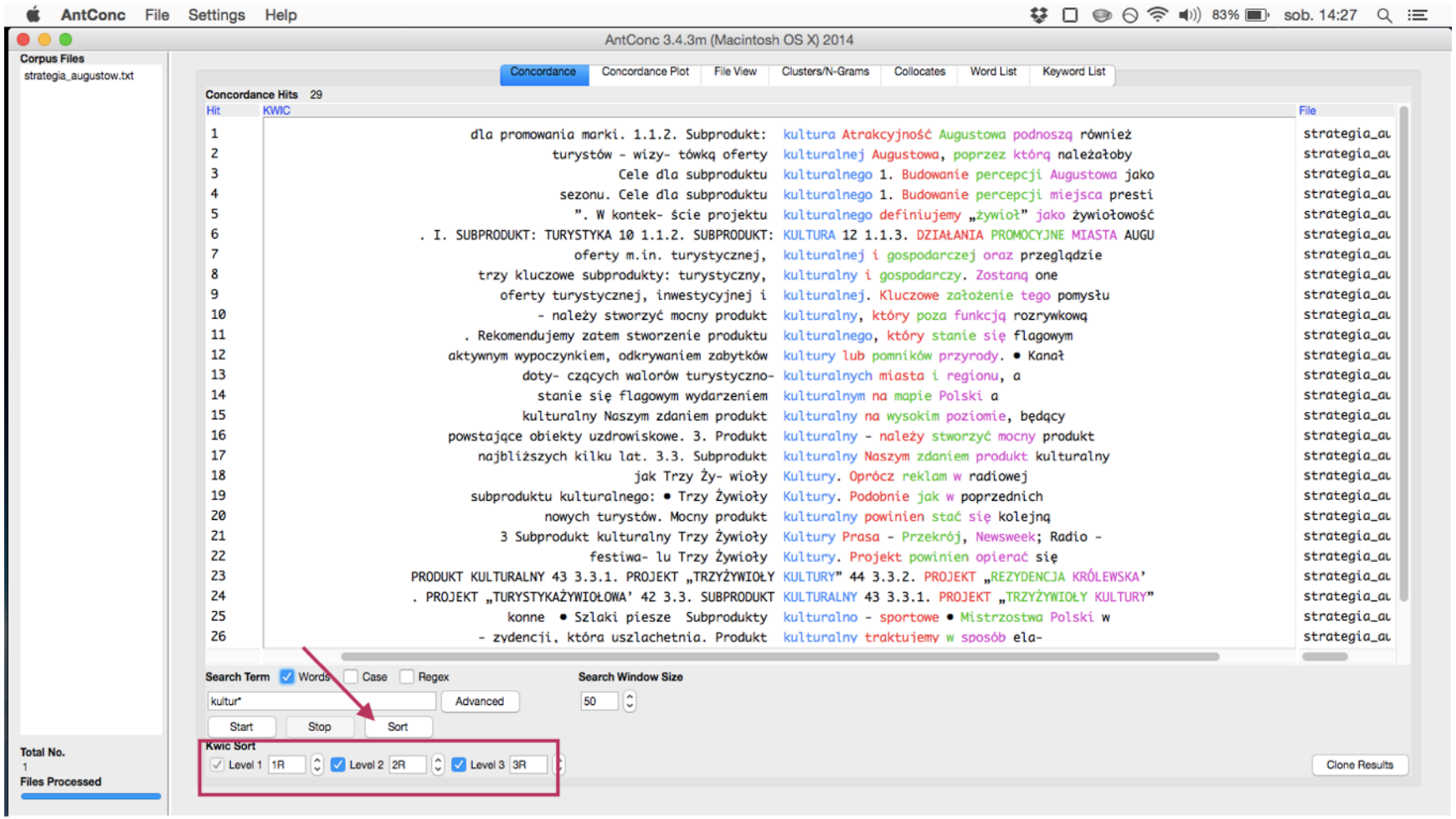
Po wpisaniu słowa - *kultura* i uruchomieniu przycisku **Start**- w głównym okienku z wynikami pojawia się wyszukiwane słowo w kontekście tekstu.



Moduł wyszukiwania nie rozróżnia wielkich i małych liter. Dodatkowo w swym działaniu program pozwala na stosowanie znaków globalnych *. Tym razem wyszukamy słowo *kultura*, a także *kulturalny*, *kulturalne*, *kulturalnego*. Umieszczenie znaku globalnego po słowie *kultura** - zadziała w taki sposób, że program wszystkie słowa zaczynające się do tego rdzenia



Dodatkowo program pozwala na sortowanie wyników alfabetycznie. Domyślnie ustawione jest sortowanie słów po prawej stronie. Na pierwszym poziomie sortowane jest pierwsze słowo po prawej, następnie drugie i wreszcie trzecie. Kolejność tą można zmienić, jak również kierunek sortowania, w taki sposób, że będą sortowane słowa po lewej stronie.



Zmiana trybu sortowania z 1R (1 prawy) - 2R (2 prawy)- 3R (3 prawy) na 1L (1 lewy) - 2L (2 lewy) - 3L (3 lewy) ujawnia ciekawszą cechę tekstu. Na 29 wystąpień słowa *kultura* - 15 razy pojawia się ono w kontekście *produktu* i *subproduktu*.