**Raport z projektu “Language Detection Comparison” na przedmiot Przetwarzanie Języka Naturalnego**

Mikołaj Wach

Kraków, 11.12.2023r.

Spis treści

1. Abstrakt 3
2. Wstęp 3
   1. Cel 3
   2. Zakres 3
   3. Metodyka 3
3. Część Teoretyczna 4
4. Część Praktyczna 6
5. Podsumowanie 13
6. Abstrakt

W poniższym raporcie przedstawiono projekt aplikacji do wykrywania języka dla podanego tekstu z wykorzystaniem biblioteki fasttext wraz z bibliotekami langdetect,polyglot i gcld3.

1. Wstęp
   1. Cel

Celem projektu jest porównanie efektywności wykrywania języka biblioteki fasttext w zestawieniu z trzema innymi bibliotekami tj. langdetect,polyglot i gcld3 oraz zaimplementowanie interfejsu graficznego.

* 1. Zakres

Biblioteki wykorzystują domyślny model danych. Testy przeprowadzone zostały na 3 językach: angielski, polski i chiński.

* 1. Metodyka

Aplikacja napisana została w języku Python. Wykorzystując biblioteki:

* tkinter i tkkbootstrap - biblioteki wykorzystane w do utworzenie graficznego interfejsu
* iso639-lang - wykorzystana do konwersji kodów języka standardu iso639-1 na ich pełne nazwy
* fasttext, langdetect, polyglot, gcld3 - biblioteki użyte do identyfikacji języka podanego tekstu

1. Część Teoretyczna

Wykrywanie języka to proces wykonywany przez komputer poprzez analizę tekstu z wykorzystaniem algorytmów NLP. Wykrywanie języka jest wykorzystywane między innymi do: rozpoznawania wiadomości w komunikatorach, bądź wiadomości email, wykorzystywane jest także w translatorach oraz przy wyszukiwaniu stron internetowych i wiele innych.

**FastText** to biblioteka realizująca rozszerzenie modelu Word2Vec, które doskonale sprawdza się w identyfikacji języka. Stworzona przez firmę Facebook. Zapewnia implementację nawet dla 294 języków, w tym projekcie zastosowany został model dla 176 języków. Cechy:

* Wydajna i dokładna identyfikacja języka
* Uwzględnia informacje o podsłowach dla lepszych wyników
* Odpowiedni dla różnych długości tekstu
* Wstępnie wytrenowane modele dostępne dla wielu języków

**Langdetect** to lekka biblioteka, która zapewnia wykrywanie języka przy minimalnej konfiguracji. Jest ona oparta na interfejsie API wykrywania języka Google. Wspiera 55 języków. Cechy:

* Lekkie i szybkie wykrywanie języka
* Oparta na analizie częstotliwości n-gramów
* Obsługuje szeroki zakres języków
* Zaprojektowany z myślą o prostocie i łatwości użytkowania

**Polyglot** to wszechstronna biblioteka, która doskonale sprawdza się w wykrywaniu języków z różnych skryptów. Obsługuje ponad 100 języków(196 dla wykrywania języków w tekście), co czyni ją cennym narzędziem dla aplikacji wielojęzycznych. Polyglot to biblioteka Pythona, która zapewnia różne funkcje NLP, w tym wykrywanie języków, rozpoznawanie nazwanych encji i wiele innych. Cechy:

* Wszechstronna wielojęzyczna biblioteka NLP
* Obsługuje wykrywanie języka, rozpoznawanie jednostek nazwanych, tagowanie części mowy i nie tylko
* Wstępnie wytrenowane modele dostępne dla różnych języków
* Zaprojektowany dla aplikacji wielojęzycznych

**Gcld3** (Google Compact Language Detector 3) to narzędzie do identyfikacji języka opracowane przez Google, które zapewnia szybkie i niezawodne wykrywanie języka dla ponad 100 języków. Zostało zaprojektowane do pracy na krótkich fragmentach tekstu i oferuje doskonałą wydajność. Cechy:

* Szybka i lekka identyfikacja języka.
* Możliwość obsługi krótkich tekstów.
* Zapewnia dokładne wyniki wykrywania języka.
* Obsługuje szeroki zakres języków.

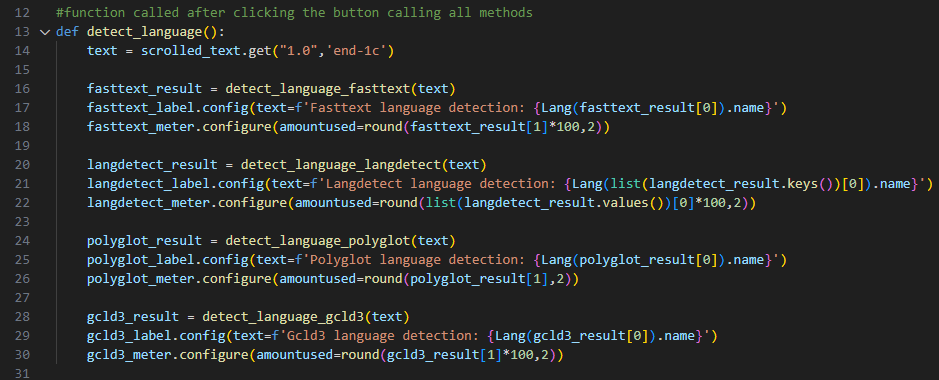
1. Część Praktyczna

Implementacja:

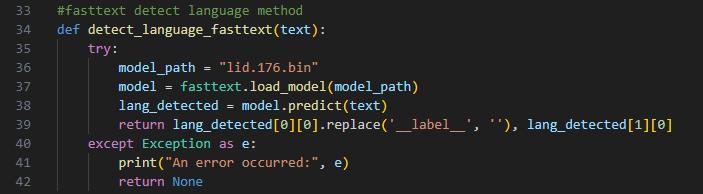
* GUI



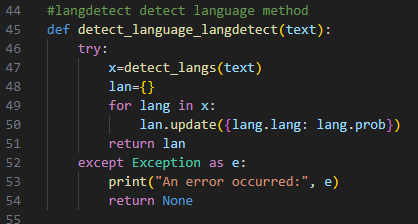
* funkcja aktywowana po kliknięciu przycisku, wywołująca funkcje odpowiedzialne z rozpoznawanie języka i podmianę danych elementów GUI



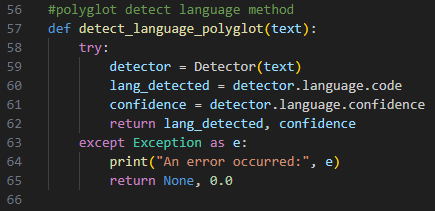
* funkcja realizująca detekcję języka przy wykorzystaniu biblioteki fasttext



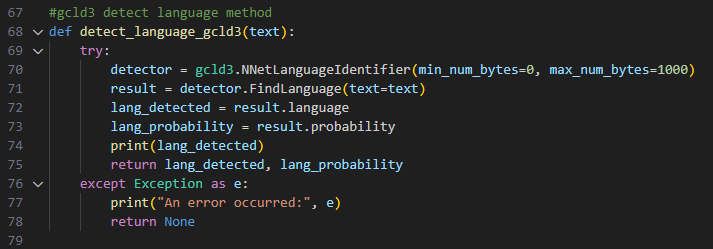
* funkcja realizująca detekcję języka przy wykorzystaniu biblioteki langdetect



* funkcja realizująca detekcję języka przy wykorzystaniu biblioteki polyglot

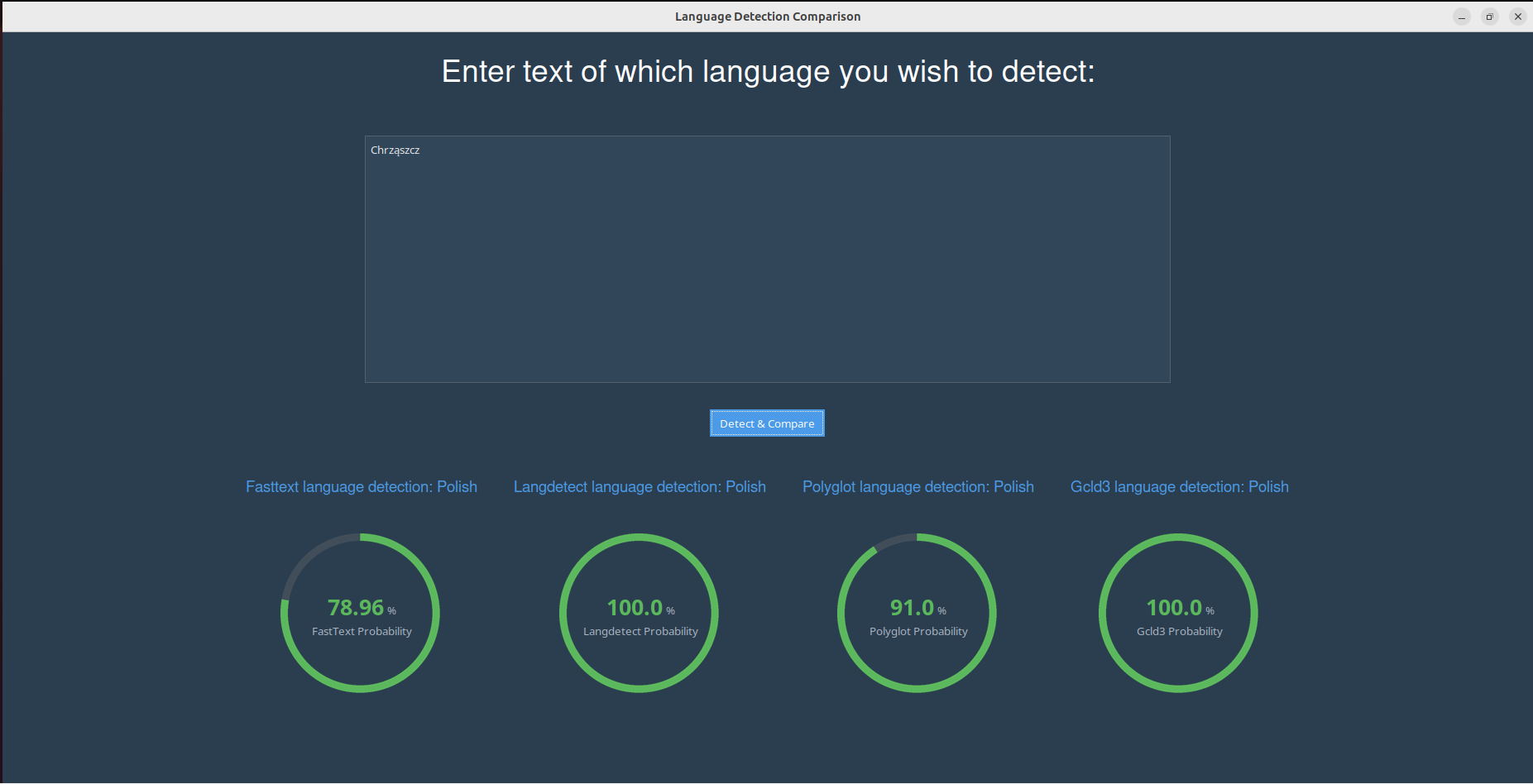


* funkcja realizująca detekcję języka przy wykorzystaniu biblioteki gcld3

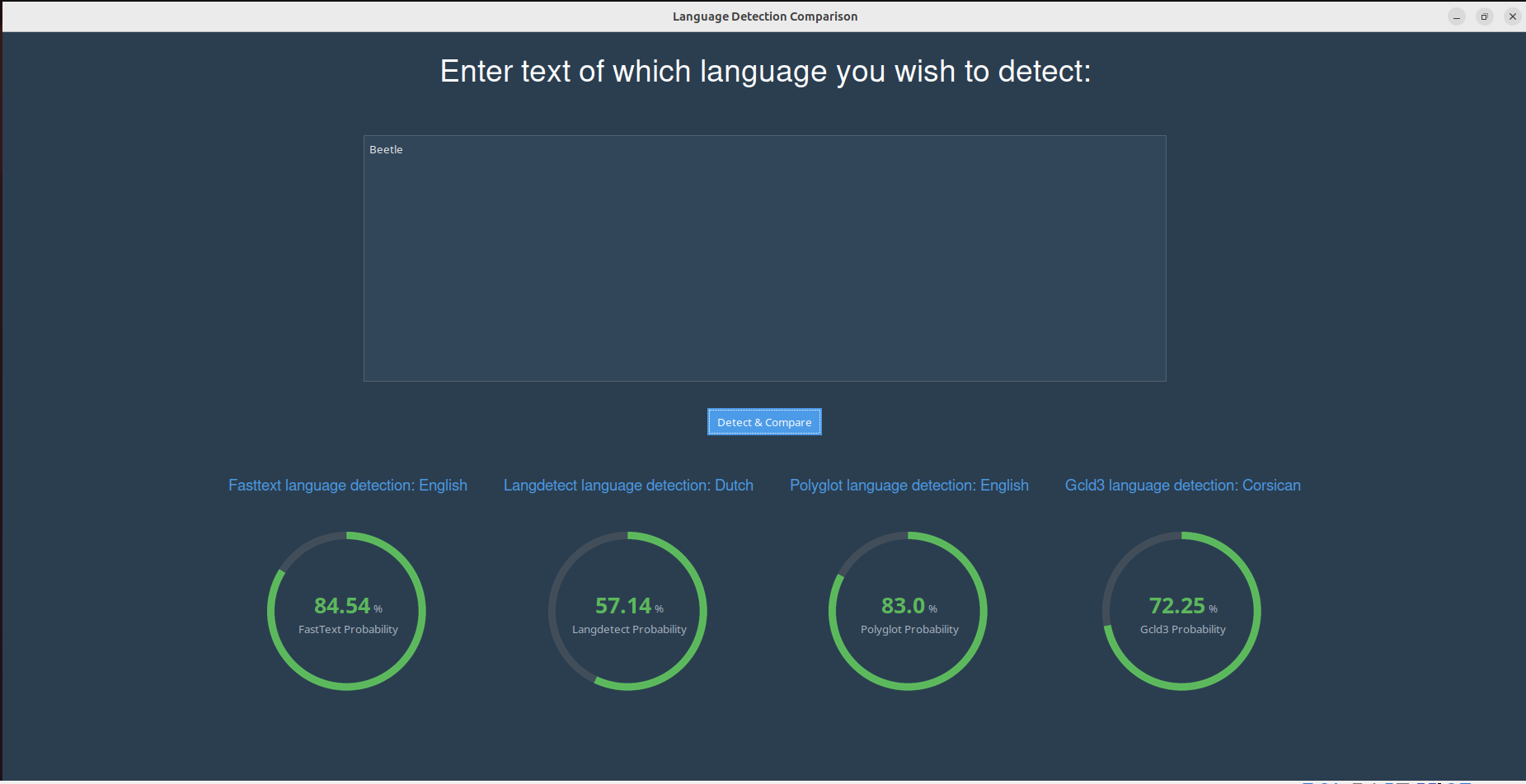


Test rozpoznawania języka dla wyrazów:

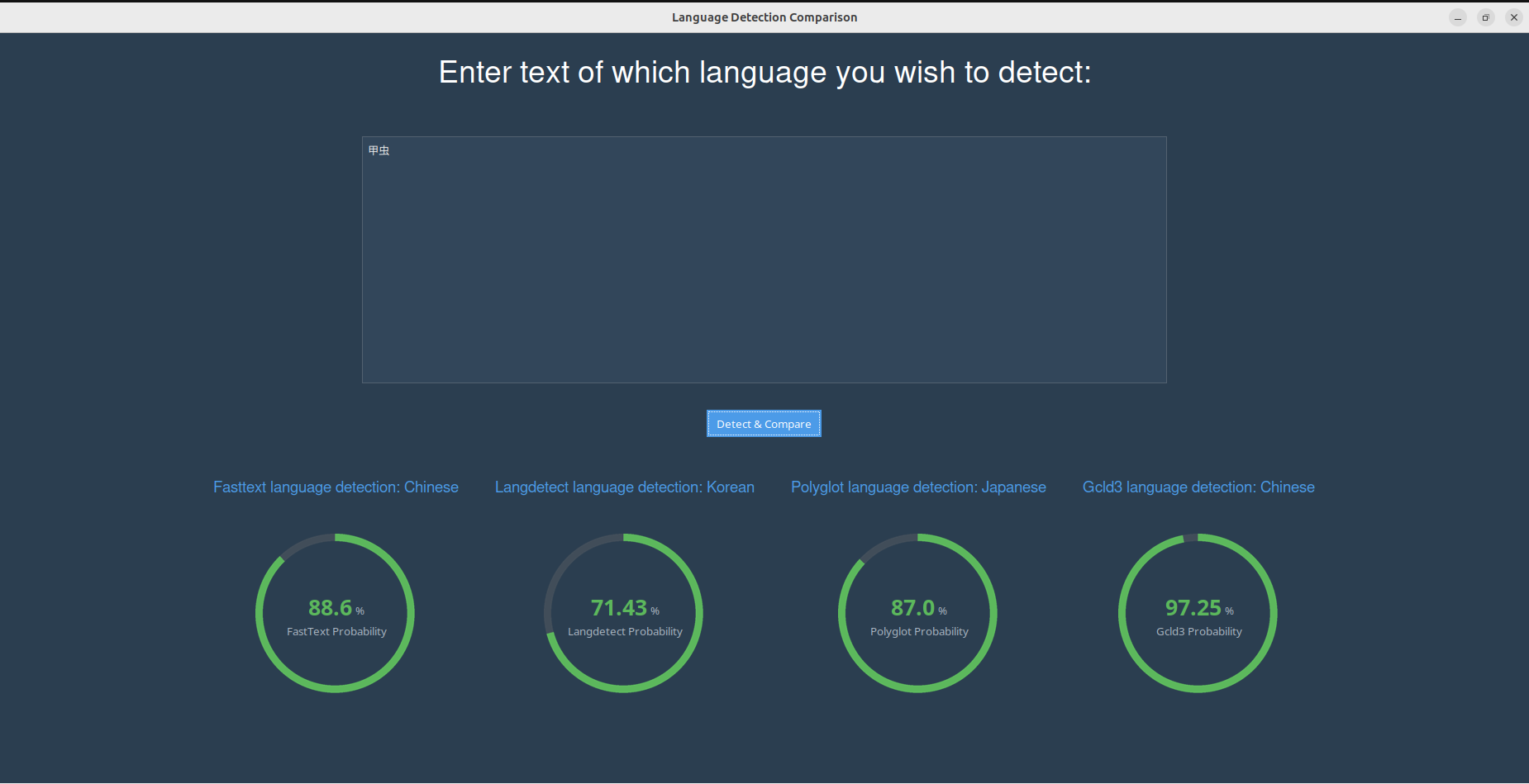
* “Chrząsz”



* Beetle

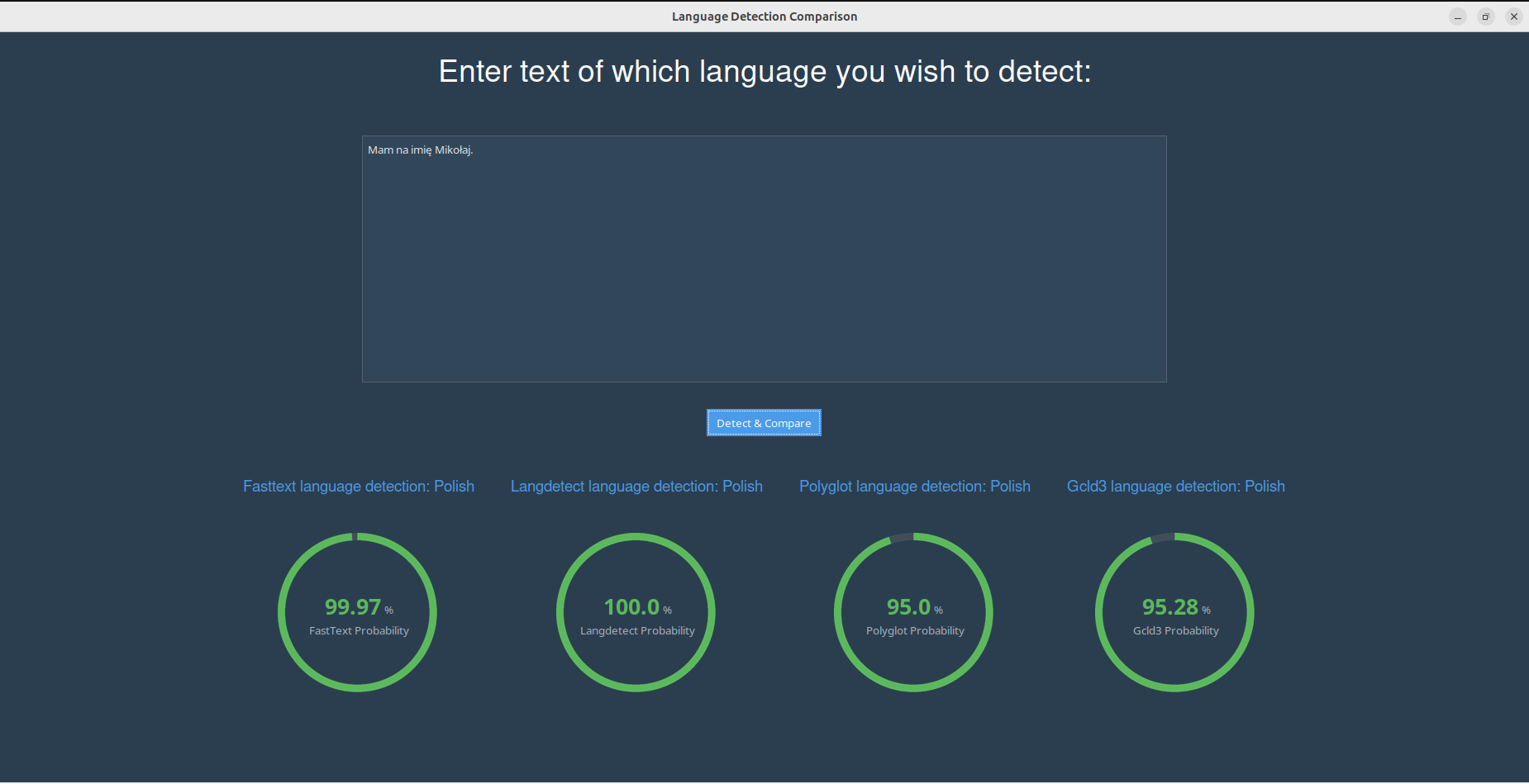


* 甲虫

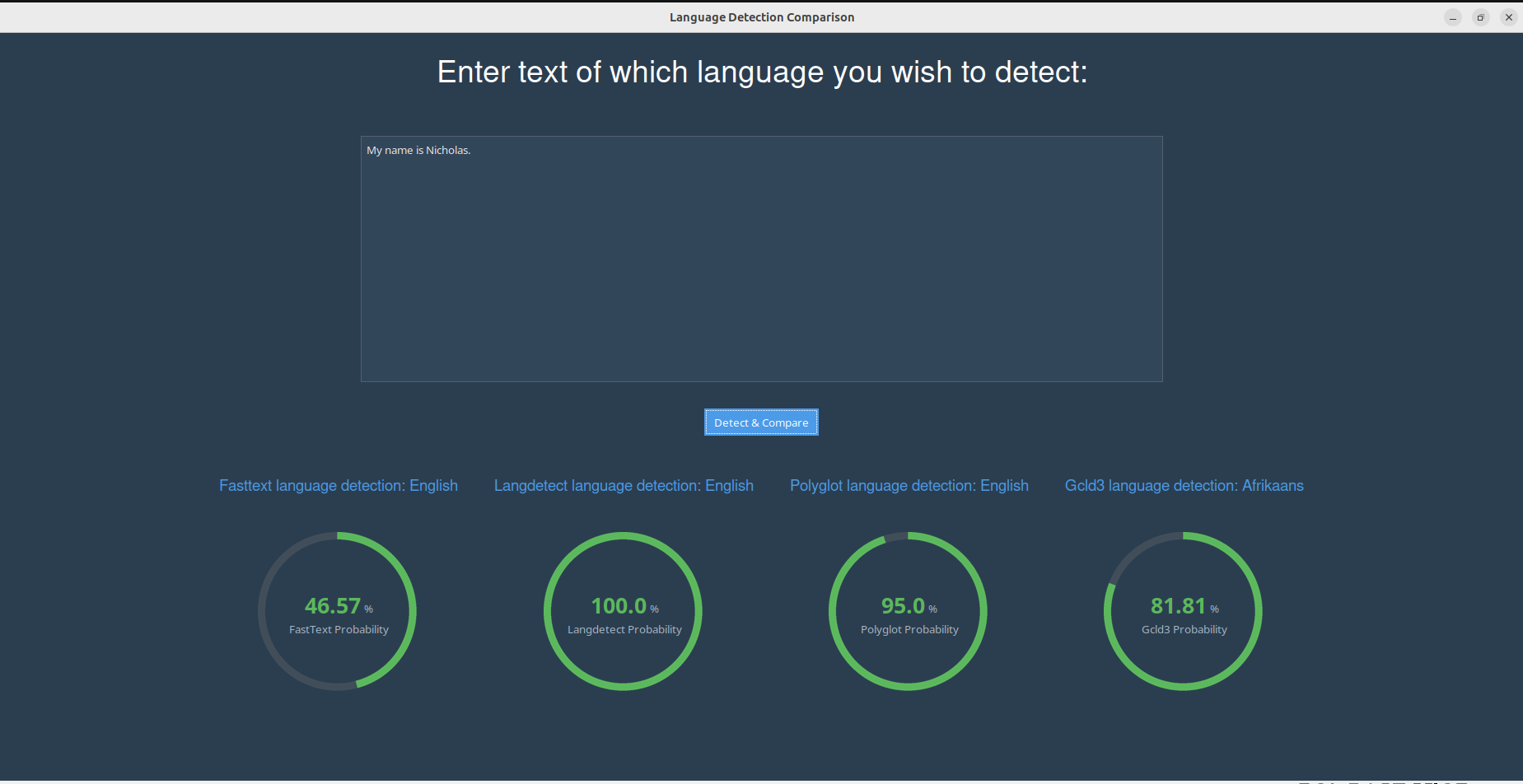


Test rozpoznawania języka dla jednego zdania:

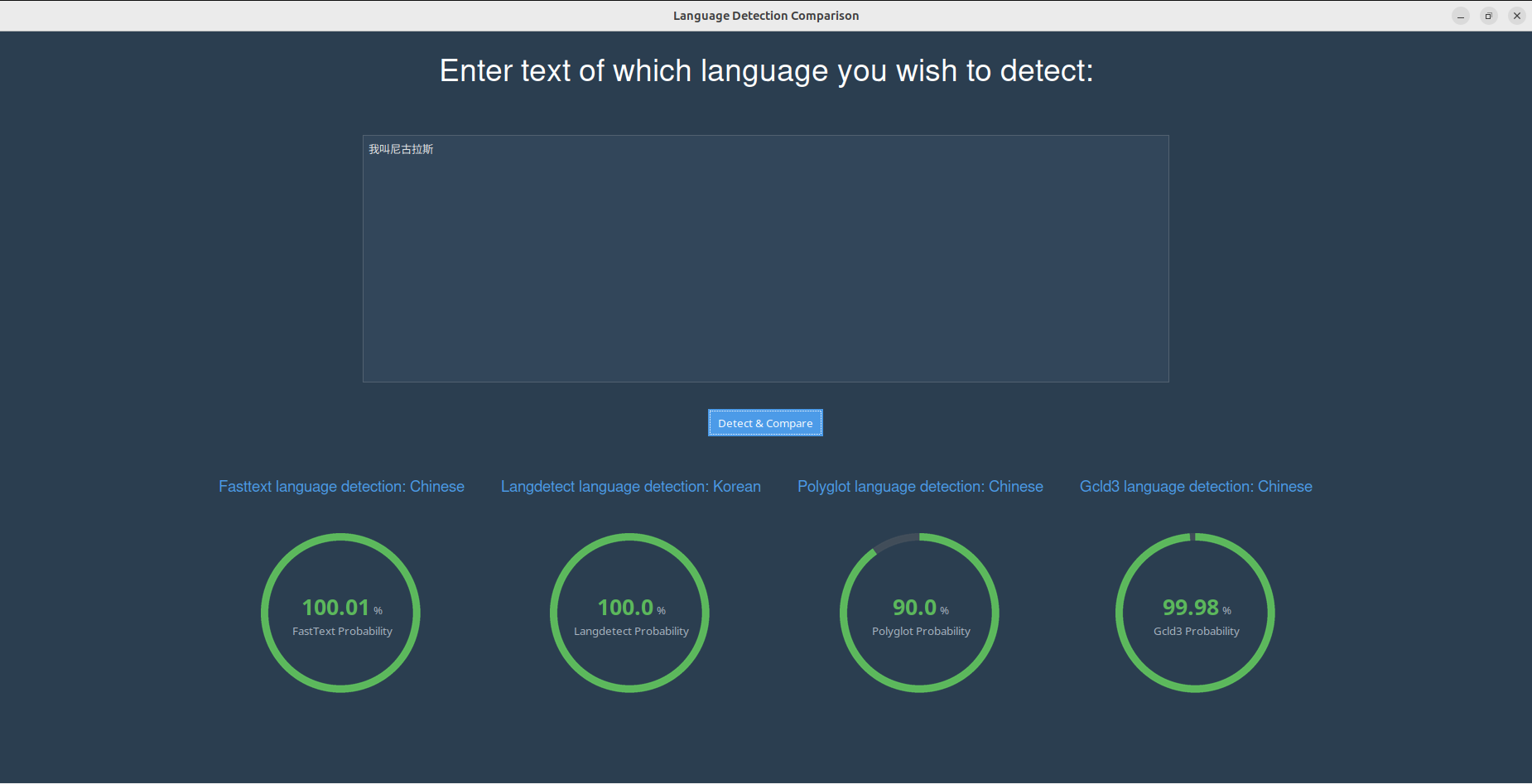
* Mam na imię Mikołaj.



* My name is Nicholas.

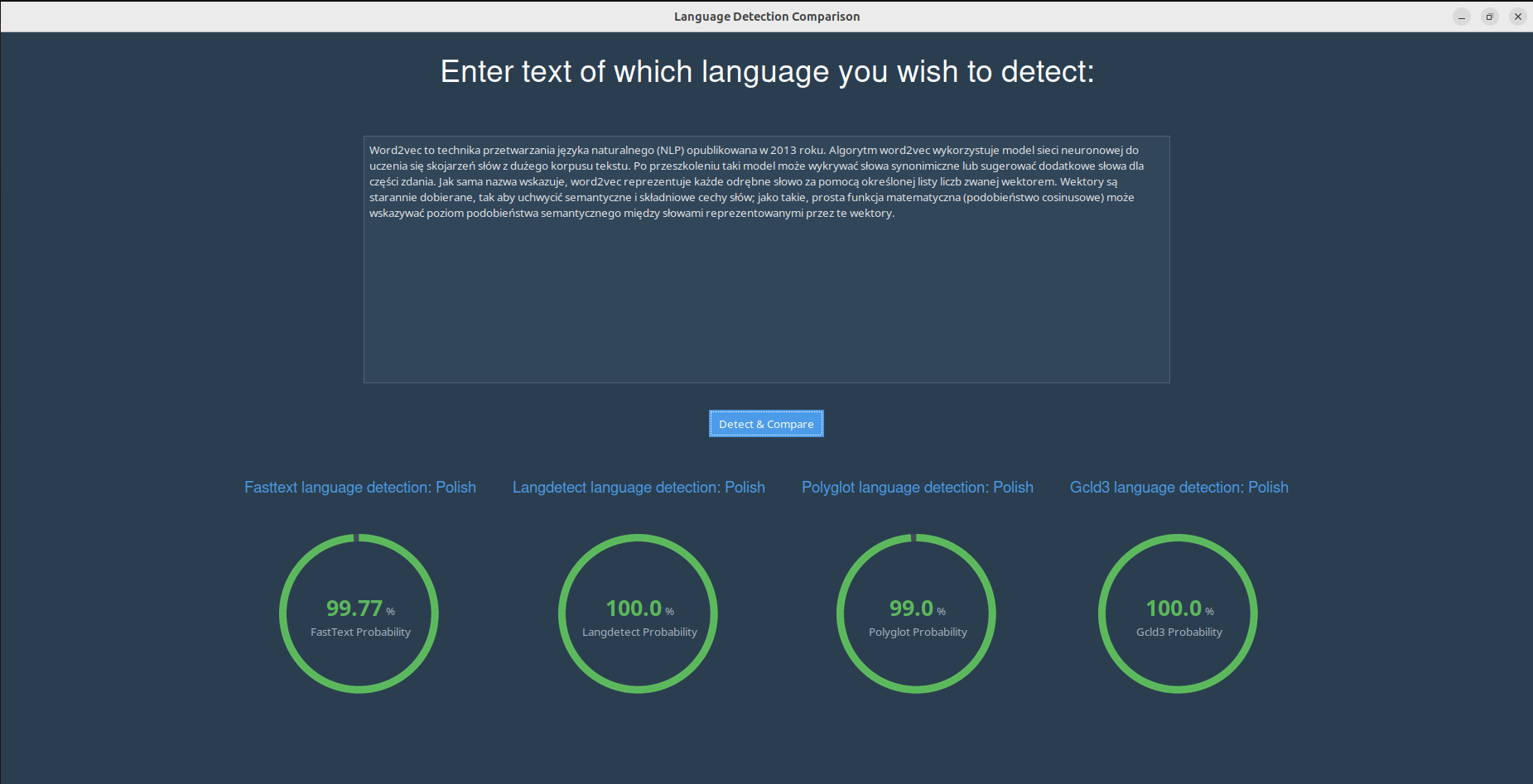


* 我叫尼古拉斯

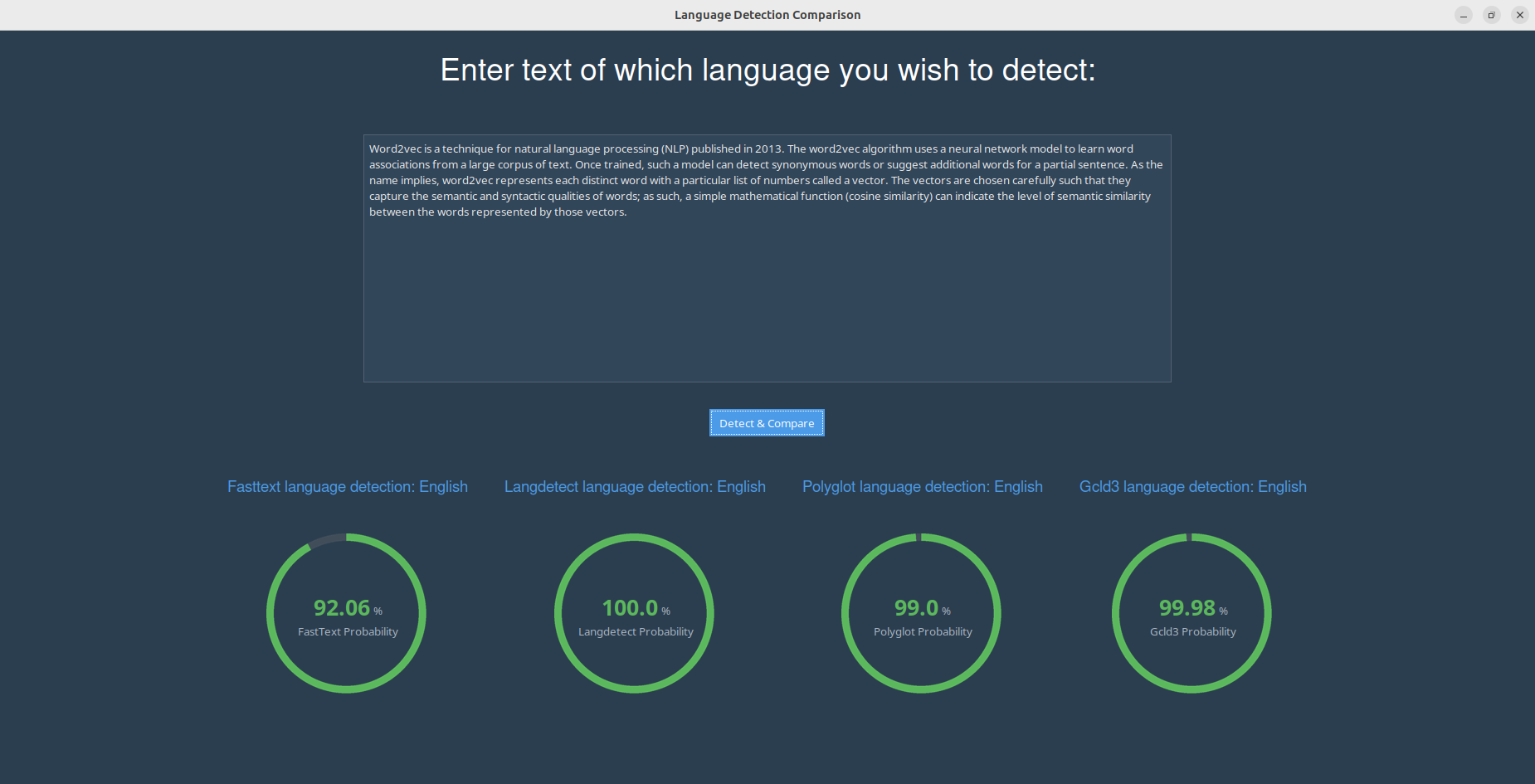


Test rozpoznawania języka dla krótkiego tekstu:

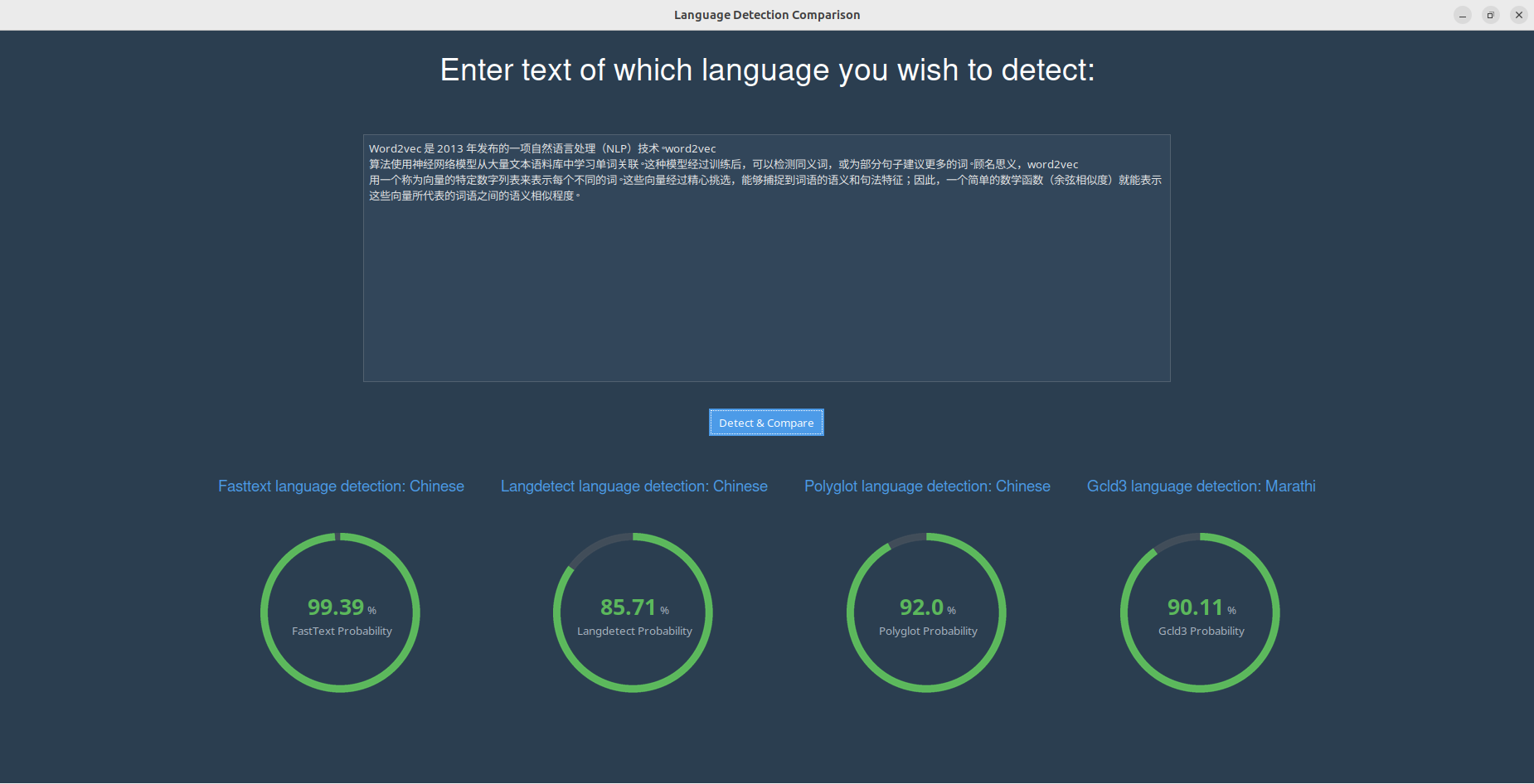
* polski



* angielski



* chiński

****

1. Podsumowanie

Z przeprowadzonych testów wynika, że biblioteka fasttext z wykorzystaniem modelu [lid.176.bin](https://dl.fbaipublicfiles.com/fasttext/supervised-models/lid.176.bin) trenowanego na danych z Wikipedia, Tatoeba i SETimes wykrywa języki z największą skutecznością w porównaniu do domyślnych modeli bibliotek langdetect,polyglot i gcld3.