Day 1 Task 4: Languages

You are to write an interactive program that, given a sequence of Wikipedia excerpts (see example below), guesses the language of each, in turn. After each guess, your program is given the correct answer, so that it may learn to make better guesses the longer it plays.

Each language is represented by a number L between 0 and 55. Each excerpt has exactly 100 symbols, represented as an array E of 100 integers between 1 and 65 535. These integers between 1 and 65 535 have been assigned arbitrarily, and do not correspond to any standard encoding.

You are to implement the procedure **excerpt(E)** where **E** is an array of 100 numbers representing a Wikipedia excerpt as described above. Your implementation must call **language(L)** once, where **L** is its guess of the language of the Wikipedia edition from which E was extracted. The grading server implements **language(L)**, which scores your guess and returns the correct language. That is, the guess was correct if **language(L)** = **L**.

The grading server calls **excerpt(E)** 10 000 times, once for each excerpt in its input file. Your implementation's *accuracy* is the fraction of excerpts for which **excerpt(E)** guessed the correct language.

You may use any method you wish to solve this problem. *Rocchio's method* is an approach that will yield accuracy of approximately 0.4. Rocchio's method computes the similarity of E to each language L seen so far, and chooses the language that is most similar. Similarity is defined as the total number of distinct symbols in E that appear anywhere amongst the previous excerpts from language L.

Note that the input data have been downloaded from real Wikipedia articles, and that there may be a few malformed characters or fragments of text. This is to be expected, and forms part of the task.

Example

For illustration only, we show the textual representation of excerpts from 56 language-specific editions of Wikipedia.

1. Yshokkie word meestal in Kanada , die noorde van die VSA en in Europa gespeel. Dit is bekend as 'n b
2. المنتج الحدي Marginal Producer ، وهو المنتج الذي يجعل المنظم لا يكسب ربحا ولا يخسر ويحصل على دخل يكف

3. "BAKILI" Futbol Klubu 1995-ci ildə Misir Səttar oğlu Əbilov tərəfindən yaradılmış və həvəskar futbol
4. Квинт Фулвий Флак (Quintus Fulvius Flaccus; † 205 пр.н.е.) е политик и генерал на Римската република
5. ইন্ডিয়ান ইনস্টিটিউট অফ সোশ্যাল ওয়েলফেয়ার অ্যান্ড বিজনেস ম্যানেজমেন্ট (সংক্ষেপে আইআইএসডব্লিউবিএম)
6. 5. juni (lipanj) (5.6.) je 156. dan godine po gregorijanskom kalendaru (157. u prestupnoj godini
7. La Caunette és un municipi francès , situat al departament de l' Erau i a la regió de Llenguadoc-Ros
8. Praha je malé městečko v Texasu , které leží cca 85 km na jihozápad od Austinu . Bylo založeno
9. Graeme Allen Brown (født 9. april 1979 i Darwin , Northern Territory , Australien) er en australsk
10. Der Plattiger Habach (3.214 m ü. A. , nach anderen Angaben nur 3.207 m [1])
11. Το Νησί Γκρέιτ Μπάρριερ (Αγγλικά : Great Barrier Island , Μαορί : Motu Aotea) είναι νησί στα βόρει
12. Sid Bernstein Presents... is a 2010 feature-length documentary film by directors Jason Ressler and E
13. El término latino lex loci celebrationis aplicado al derecho internacional privado quiere decir: "le
14. Apollo 5 oli kosmoselaev , mis sooritas Apollo programmi teise mehitamata lennu. Lennu käigus testit
15. سیارک ۱۳۵۰ (1350 : به انگلیسی Rosselia ، 1934: نامگذاریTA) هزار و سیصد و پنجاهمین سیارک کشف شده‌اس
16. V. I. Beretti (myös Vikenty Ivanovitš Beretti , alk. Vincent Beretti ; 1781 Milano Italia – 18. elok
17. Le 5 e bataillon de parachutistes vietnamiens (ou 5 e BPVN ou encore 5 e Bawouan) est une unité par
18. Amina Sarauniyar Zazzau,, wadda ta rayu daga shekarar 1533 zuwa 1610, ɗaya ce daga cikin ‘ya'ya biyu
19. ב מתמטיקה , השערת רימן היא השערה שהציע בשנת 1859 ה מתמטיקאי ברנרד רימן , מגדולי המתמטיקאים של אותה ע
20. Sudski proces Doe protiv Boltona je sudski proces iz 1973 . godine kojim je američki Vrhovni sud uki
21. Owen Cunningham Wilson (1968 . november 18. , Dallas , Texas , Egyesült Államok) amerikai színész
22. Հայ Կաթողիկե Եկեղեցին պատկանում է Արևելյան Կաթոլիկ Եկեղեցիներին և այսպիսով ենթարկվում է Հռոմի Պապի ա
23. Dionysios dari Halicarnassus (Bahasa Yunani : Διονύσιος Ἀλεξάνδρου Ἁλικαρνᾱσσεύς , Dionysios putra
24. Nnamdi "Zik" Azikiwe , bu onye isi-ala izizi Nijiria nwere. Ochichi ya bidolu na afo 1954 welu ruo n
25. La Riserva naturale orientata Serre della Pizzuta è un'area protetta del dipartimento Regionale di S

26. 石橋和義 （いしばし かずよし/まさよし、生没年不詳）は、 詳。 石橋氏 初代当主。初名氏義。 尾張 三郎を通称とし、官途は、 左近将監 → 三河守 → 左衛門佐 。 足利直義
27. კორბინ ბლიუ (ინგლ. Corbin Bleu ; დ. 21 თებერვალი , 1989 , დაბადების ადგილი ბრუკლინი , ნიუ-იორკი , ა
28. Та́рья Ка́арина Ха́лонен (Tarja Kaarina Halonen)); 24 желтоқсан , 1943 , Каллио , Хельсинки , Финлан
29. 딜롱 (Dilong)은 중국 랴오닝(Liaoning) 지방의 익시안층(Yixian Formation)에서 온전한 4구의 화석으로 발견되었다. 이 공룡은 가장 원시적인 초기의 티
30. Сүймөнкул Чокморов - советтик актёр. Жетинин айынын 9 (ноябрь) 1939-жылы, Фрунзе шаарын жанындагы Чо
31. D' Mirjam vun Abellin war eng Nonn a Mystikerin , och " Maria vum gekräizegte Jesus " genannt. Si as
32. Panopea abrupta (angl. Geoduck) - jūrinių dvigeldžių moliuskų rūšis, priklausanti Hiatellidae šeim
33. "Dzimis Latvijā" ir Liepājas dueta Fomins & Kleins 2004 . gada 23. februārī izdotais otrais albu
34. I Ludwik Lejzer Zamenhof dia dokotera mpijery maso nipetraka any Polonia . Fantantsika izy ankehitri
35. Седумстотини милиони малечки алвеоли во белите дробови , всушност се шупливи чаури - алвеоли прекрие
36. Энэхүү шувуу нь Бутан , Хятад , Гонконг , Энэтхэг , Пакистан , Иран , Япон , Казакстан , Солонгос ,
37. भारतातील महाराष्ट्र राज्याच्या नागपूर पासुन २१६ कि.मी. दूर असलेले एक गाव. ते वैनगंगा नदीच्या काठावर
38. De Slotervaart was oorspronkelijk de waterweg die sinds de Middeleeuwen het dorp Sloten verbond met
39. Macierz S (macierz rozpraszania, od ang. scattering matrix) jest centralnym elementem w mechanice k
40. A Hora do Rush 3 (Rush Hour 3 , no original) é o terceiro filme da franquia Rush Hour . Dirigido po
41. Coordonate : 51°34′0″N 12°3′0″E / 51.56667 , 12.05 Brachstedt este o comună din landul Saxonia-A
42. Гробницы императоров династии Мин и Цин — памятник Всемирного наследия ЮНЕСКО , состоящий из несколь
43. Kovalentni radijus atoma - ponekad se naziva i valentni radijus. Kovalentni radijus je srednje rasto
44. Koniecpol je mesto v Poľsku v Sliezskom vojvodstve v okrese Powiat częstochowski v rovnomennej gmine
45. Hoxhë Vokrri vije nga Shqipëria ishte një klerik shqiptar i cili luftonte për Çështjën Kombëtare . A
46. Гурдијеље је насеље у општини Тутин у Рашком округу . Према попису из 2002. било је 93 становника (п

47. Underhållsstöd betalas ut av Försäkringskassan (FK) till en förälder som är vårdnadshavare och bor e
48. இந்தியாவின் தேசிய நெடுஞ்சாலைகள் நடுவண் அரசின் தேசிய நெடுஞ்சாலைத் துறையால் பராமரிக்கப்படுகின்றன. பெரு
49. Дар он зиндаги .маишат ,фаолияти мехнати,муборизаи ичтимои, русуму омол, хислат ва эхсосоти халк ифо
50. ไฟทอฟธอรา อินเฟสทันส (อังกฤษ : Phytophthora infestans) คือเชื้อ ราโอโอไมซีท หรือ ราน้ำที่เป็นสาเห
51. ABUL FAWARIS BERRANY - 11. asyrda Orta Aziýadaky oguz taýpalarynyň berrany dinastiýasynyň wekili. Ol
52. Egemenlik ya da hâkimiyet , bir toprak parçası ya da mekan üzerindeki kural koyma gücü ve hukuk yara
53. Темне фентезі (від англ. Dark Fantasy - темне, похмуре фентезі) - піджанр літератури, який включає
54. Paris By Night 84: In Atlanta - Passport to Music & Fashion (Âm nhạc và Thời trang) là chương tr
55. ISO 3166-2:GU ni akoole ninu ISO 3166-2 , apa opagun ISO 3166 ti International Organization for Stan
56. 下卡姆斯克 （俄文 ： Нижнека́мск ； 韃靼語 ： Түбән Кама/Tübän Kama ）是 俄羅斯 韃靼斯坦共和國 東北部的一個城市，位於 卡馬河 南岸。 2002年 人口22

The sample input file `grader.in.1` contains 10 000 such examples. The 56 languages are those listed as "mother tongue" in the IOI 2010 registration data. The language for each excerpt is chosen at random from these 56 languages, and each excerpt is taken from the first paragraph of an article chosen at random from the corresponding Wikipedia edition. Each line of the file contains:

- The two-letter ISO code for the Wikipedia language edition;
- 100 numbers between 1 and 65 535, representing the first 100 symbols, in sequence, of the first paragraph of the article;
- a viewable representation (in UTF-8) of the 100 symbols that you can read in your text editor or Firefox web browser. This viewable representation is for your convenience only, and is not intended to be used as input for your program.

The official grader uses 10 000 different excerpts, selected in the same way from the same 56 Wikipedia editions. However, the grader assigns a different number between 0 and 55 to each language, and a different number between 1 and 65 535 to each symbol.

Subtask 1 [30 points]

Your submission must achieve accuracy of 0.3 or better on the grading server.

Subtask 2 [up to 80 points]

Your score will be 114(α-0.3), rounded to the nearest integer, where α is the accuracy of your submission.

Implementation Details

- Implementation folder: `/home/ioi2010-contestant/language/`
- To be implemented by contestant: `lang.c` or `lang.cpp` or `lang.pas`
- Contestant interface: `lang.h` or `lang.pas`
- Grader interface: `grader.h` or `graderlib.pas`
- Sample grader: `grader.c` or `grader.cpp` or `grader.pas` *and* `graderlib.pas`
- Sample grader input: `grader.in.1`.
  Note: Each line of input contains: a two-character language code; an excerpt represented as 100 numbers separated by spaces; the text representation of the excerpt.
- Expected output for sample grader input: *If the implementation calls language as specified for each of the 10 000 examples, the sample grader will output* `OK alpha` *where alpha is the accuracy.*
- Compile and run (command line): `runc grader.c` or `runc grader.cpp` or `runc grader.pas`
- Compile and run (gedit plugin): *Control-R*, while editing any implementation file.
- Submit (command line): `submit grader.c` or `submit grader.cpp` or `submit grader.pas`
- Submit (gedit plugin): *Control-J*, while editing any implementation or grader file.