Linguistic Hacking
How to know what a text in an unknown language is about?

Martin.Haase@uni-bamberg.de

24th Chaos Communication Congress

It is sometimes necessary to know what a text is about, even it is written in a language you don't know. This can be quite problematic, especially if you do not even know in what language it is written. This talk will show how it is possible to identify the language of a written text and get at least some information about the contents, in order to decide whether a specialist and which specialist is needed to know more.

1 Introduction

In a first and rather brief outline, I will show how to identify the language of a written text in traditional ways and with the help of computer technology. In the second part, I will show how to get at least some information out of an unknown text. This is all about linguistics, but what has it to do with hacking? I will show that some tricks must be used to solve such problems and define hacking in this context according to Eric Raymond's seventh definition as "the intellectual challenge of creatively overcoming or circumventing limitations." [11, 234]

I will confine my analysis to written texts (not necessarily in Roman script), although, based on a multi-language corpus of telephone calls [8], considerable progress has been made in the identification of spoken languages [9]. The main reason for this omission is that with spoken language it is far more difficult (and perhaps even impossible) to get clues about the contents of a conversation without at least some knowledge of the language in question.

2 How to identify a language

2.1 The traditional approach

If the text comes in a non-Roman and non-Cyrillic writing system, it is in most cases quite easy to identify the script and the language, because exotic scripts are often language-

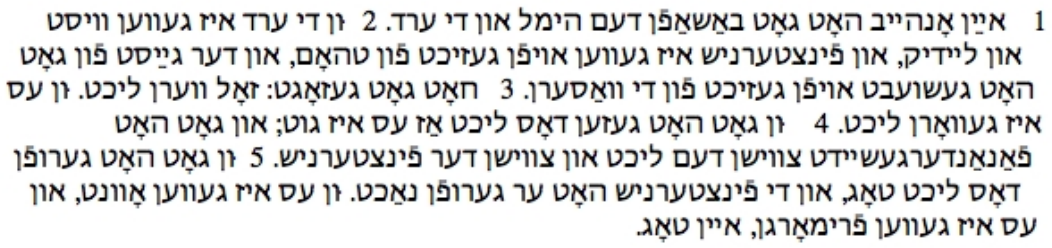

Figure 1: Beginning of Genesis in Yiddish

specific. A handbook on writing systems [4] or web resources [1] can easily help to identify a script and thereby the language.

There are some difficult cases of course. One such case is the Hebrew script which is used for:

- Old and Modern Hebrew,
- Ladino (with different varieties),
- Judeo-Arabic,
- Yiddish

Of course, there are some simple tricks to distinguish between Hebrew and the other languages. Normally, Hebrew is written without vowel diacritics (the little dots over and under Hebrew letters). If your text shows no such signs, it is probably Hebrew. If it contains such "vocalization signs", it may still be Hebrew (a text from the Bible, from a children's book, or from learning material), but in that case the vocalization can be consistently found throughout the text. If some words show (some) vocalization and others don't, it is most probably a Yiddish text, where Yiddish words contain a subset of vocalization signs, but loan words from Hebrew are used without vocalization. Ladino doesn't contain super- or subscript diacritics at all. Moreover, Yiddish and Ladino texts may contain Roman-script arabic numbers and Roman-script punctuation signs, but sometimes even Hebrew texts contain western numbers. Figure 1 shows a Yiddish text (few vocalization, Roman-script arabic numbers, Western punctuation), whereas figure 2 shows the same text from the Hebrew bible (with full vocalization), i. e. the beginning of Genesis, the first book of the Bible (Hebrew numbering, full vocalization, non-Western punctuation).

The problem gets worse when we turn to the Arabic writing systems. Variants are used for about twenty different and partly unrelated languages (and more subvarieties) and Modern Arabic itself has about thirty commonly used varieties. In order to get an idea about the language, it is helpful to work with sample texts [1, 5, 7].

The Cyrillic writing system is even worse, since it is used for more than sixty languages. Cyrillic writing systems for non-slavic languages were conceived mainly in the

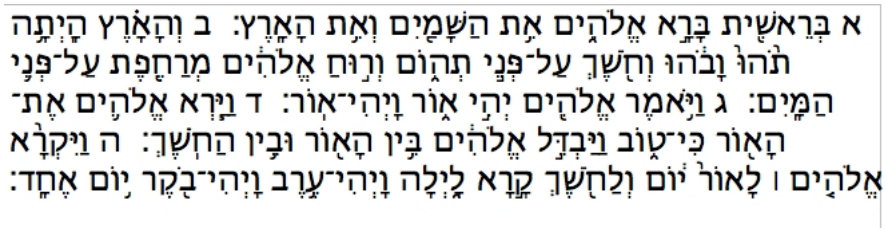

Figure 2: Beginning of Genesis in Biblical Hebrew

middle of the 20th century. When Cyrillic was adapted to different phonological systems, additional letters were introduced that make it easy to identify a language, because every writing system contains different special signs. That is why the identification of Cyrillic languages is mainly done through the identification of character encoding.

2.2 Computer-aided language identification

There are three common techniques [12]:

1. frequencies of unique characters and character strings: this method, known from cryptoanalysis, classifies documents by the frequency of unique characters and the occurrence of typical character strings; a nifty variant of this approach consists in measuring the compression efficiency that a program such as *gzip* achieves when appending an unknown document to various reference documents. [3]

2. common words recognition: this method is based on word frequency lists (generated from sample texts), the unknown text is analyzed word by word and compared to the list of the top 100 words (or so) of the sample texts;

3. n-gram analysis: this method works like common words recognition with the difference that (instead of words) sequences of n characters are used (2-character sequences, 3-character sequences, etc.): if we split the word *text* into 3-grams, this would be the result: `(_TE)`, `(TEX)`, `(EXT)`, `(XT_)`, _ denoting the word boundary.

These approaches all work according to the scheme in Figure 3: a document model is generated from the input text in the unknown language and then this model is compared to the existing models generated from sample texts.

The advantages and shortcomings of this procedure can be critically evaluated [6]: the main drawbacks are that only a closed class of languages can be identified (dialects and varieties of these languages are usually ignored), and normally, multilingual text cannot be processed. If the programs work for non-Roman scripts, they usually reduce the recognition of non-Roman script languages to the detection of the encoding which doesn't work if a writing system is used for several languages and if non-standard or mixed character encodings are used.

Here is a list of free software readily available (and running) on the internet [6, 13, 14]:

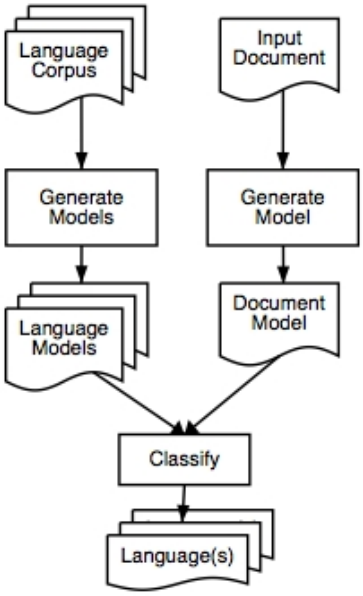

Figure 3: Language Identification Workflow [10]

- TextCat (`http://odur.let.rug.nl/vannoord/TextCat/Demo/`), an n-gram based identification tool for 76 languages, usable as a web application,
- Languid (`http://languid.cantbedone.org/`), a downloadable program, the web application is not running properly,
- Langid (`http://complingone.georgetown.edu/~langid/`), a web-based identification tool for 65 languages, based on n-gram analysis,
- LanguageGuesser (`http://www.xrce.xerox.com/cgi-bin/mltt/LanguageGuesser`) provides for the web-based identification of about 40 languages, based on statistical methods (frequency tests on characters and character sequences) [2],
- Polyglot 3000 (`http://www.polyglot3000.com/`), closed-source Windows freeware, identifying currently 441 languages, corpora and method are unknown.

3 How to get an idea about the contents of a text?

When we have identified the language of the text, it would be helpful to get an idea of its contents before we try and find a specialist who can help us with the translation. Perhaps the text is not interesting at all or has been translated before.

A hacker's approach to this task could be as follows:

- look for things you recognize without any help: numbers, dates, words from another language; a number or a date can be a good hint; if it is a precise number or date, a quick look-up with your preferred search engine might be helpful,
- look for typographic hints to important content: bold or italic print, colored or underlined text chunks, capital letters (they may indicate names that you may recognize or look up in Wikipedia).

Even with these steps you can get important hints about the contents of the text.

Moreover, the principle of least effort or Zipf's law [15] can be very helpful to find out what a text is about: Very frequent words are shorter and contain less lexical information, whereas infrequent words are longer and contain more lexical information; moreover, less lexical information implies more grammatical information and vice versa. For our purpose, we are looking for words with more specific lexical information. So we can ignore all short words, even if they reiterate throughout the text. A longer word that is repeated is therefore more interesting.

Here is an example (from Samoan, which is difficult to identify as such, since it is not contained in typical language sample collections):

Ua salalau lenei *gagana* i le lalolagi atoa. 'O lenei fo'i *gagana*, 'ua 'avea ma *gagana* lona lua a le tele o tagata 'o le vasa Pasefika, e pei 'o Samoa. E iai le manatu, 'o le *gagana* **fa'aperetania**, 'ua matuā talitonu i ai le tele o tagata Samoa e fa'apea 'o le *gagana* e maua ai le atamai ma le poto. 'E talitonu fo'i nisi o i latou, 'e lē aoga la latou *gagana*. E lē sa'o lea tāofi, 'auā e 'avatu le *gagana* **fa'aperetania** i Samoa, 'ua leva ona atamamai ma popoto tagata Samoa e fai lo latou soifua ma lo latou lalolagi.

The interesting words in this text are *gagana* and **fa'aperetania**, perhaps latou too, although this is short enough to be a more grammatical item. It is difficult to find a Samoan dictionary, but a quick search reveals that **fa'aperetania** means 'English' (8th Google result) and *gagana* 'language' (11th & 13th Google hit); latou is more difficult to find and less useful, since it is a third person plural pronoun (as the French Wiktionary reveals). So the text is about the English language, probably in Samoa ("*gagana* **fa'aperetania** i Samoa").

The example shows that it is rather simple to get at least minimal information out of a text whose language is unknown to us, even if we don't have direct access to a translator or a dictionary.

References

[1] Omniglot. Writing Systems and Languages of the World. `http://www.omniglot.com/` (2007-11-16).

[2] K.R. Beesley. Language identifier: A computer program for automatic natural-language identification of on-line text. *Language at Crossroads: Proceedings of the 29th Annual Conference of the American Translators Association*, pages 12–16, 1988.

[3] D. Benedetto, E. Caglioti, and V. Loreto. Language Trees and Zipping. *Physical Review Letters*, 88(4):48702, 2002.

[4] P.T. Daniels and W. Bright. *The world's writing systems*. New York etc.: Oxford University Press, 1996.

[5] K. Faulmann. *Das Buch der Schrift: Enthaltend die Schriftzeichen und Alphabete aller Zeiten und aller Völker des Erdkreises*. KK Hof-und Staatsdruckerei, 1880.

[6] B. Hughes, T. Baldwin, S. Bird, J. Nicholson, and A. MacKinlay. Reconsidering Language Identification for Written Language Resources. *eprints: http://eprints.infodiv.unimelb.edu.au/archive/00001744 (2007-11-16)*.

[7] N.C. Ingle. *Language Identification Table.* London: Technical Translation International, 1980.

[8] Y.K. Muthusamy, R.A. Cole, and B.T. Oshika. The OGI multi-language telephone speech corpus. *Proceedings of the International Conference on Spoken Language Processing*, pages 895–898, 1992.

[9] Y.K. Muthusamy and A.L. Spitz. Automatic language identification. *Cambridge Studies In Natural Language Processing Series*, pages 273–276, 1997.

[10] A. Poutsma. Applying Monte Carlo Techniques to Language Identification. *Language and Computers*, 45(1):179–189, 2002.

[11] E.S. Raymond. *The New Hacker's Dictionary*. Cambridge, Mass.: MIT Press, 1996.

[12] C. Souter, G. Churcher, J. Hayes, J. Hughes, and S. Johnson. Natural Language Identification Using Corpus-Based Models. *Hermes Journal of Linguistics*, 13(S 183):203, 1994.

[13] G. van Noorden. Language Identification Tools. `http://www.let.rug.nl/~vannoord/TextCat/competitors.html` (2007-11-16).

[14] Wikipedia. Language Identification. `http://en.wikipedia.org/w/index.php?title=Language_identification&oldid=139087517`.

[15] G.K. Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. New York: Hafner, 1965.