CHOAS THEORY AND FRACTAL DYNAMICS

TERM PROJECT:

The Dynamics of Machine Translation

Joel Ironstone

960100290

2002-05-06

Table of Contents

1 INTRODUCTION

The internet provides a medium through which people from all over the world can interact and communicate. People can exchange documents instantaneously and easily publish their ideas for everyone to discuss and enjoy. Unfortunately, this facility marginalizes people who are not able to read the languages commonly published on the internet. Several organizations and companies provide automatic translation services for web pages and text. Although these services claim to be extremely accurate, they often produce ridiculous translations that do not reflect the original text.

This paper serves as a preliminary investigation into the dynamics of these translations. Experiments are performed using the Babelfish[tm] translation software and a series of programs both developed by the author and available freely on the internet. The Babelfish[tm] translation system is analyzed by modeling it as an iterative dynamical system in which a single iteration consists of a translation from English to another language and back again. The results of these iteration cycle can be analyzed using the tools provided by the author and the theory of non-linear dynamical systems.

Most of the effort spent on improving translation systems is focused on improving the fidelity of the translations themselves. Human translators are employed to asses and score the fidelity of the machine translation and suggest improvements and discover weakness. It is postulated by the author that stable sentences (those that translate from English to another language and back onto themselves) are those sentences that are most easily translated, and that stable translation systems are the ones which provide the best results.

By performing an analysis such as the one discussed in this paper two useful ends can be met. The first being the ability to asses translation systems extensively without employing human translators. The second is the ability to pre-process text to be translated and provide suggestions of stable replacement phrases fro those may cause erroneous results after translation. The author can the tune his/her work to make it more easily translatable, and stability checks on documents can be performed just as easily as spelling and grammar checks on documents that are intended to be presented to a multilingual audience.

2 MATHEMATICAL BACKGROUND

2.1 Topology of the Set of English Sentences

There exists a discrete set containing all valid English sentences. This set has infinite cardinality and can thus be partitioned in infinite ways. The set of English sentences is a subset of the set of all ordered collections of English words. The project will consider finite subsets of the set of English sentences. The elements of these chosen subsets will be used as initial conditions in analyzing the dynamical mapping of the computer translation software. According to the results of these mappings, the subsets described will be partitioned into qualitative groups as listed in 2.2.

2.1.1 The Set of English Sentences as a Metric Space

A metric space is a space **X,** together with a function **d:X x X-> R** that has the following properties.

1) $d(X,Y)=d(Y,X)$
2) $0<d(x,y)<\infty$ $X \neq Y$
3) $d(x,x)=0$
4) $d(x,y) \leq d(x,z)+d(z,y)$

(Barnsley 11)

To analyze trajectories in the way described in section 2.2, it is important to develop a metric that assesses distances between points in the space of English sentences, the author has developed such a metric, and it is described in section 4 of this paper.

2.2 Autonomous Systems Theory

Autonomous Systems Theory presents us with a framework for understanding the behaviour of dynamical systems under different initial conditions. The iterative experiments described in this paper produce trajectories in the state space of English sentences. Points in this space (unique sentences) can be mapped to other sentences, themselves, or outside of the space of English sentences to the space of collections of infinite words. By qualitatively analyzing these trajectories, the behaviour of the system can be categorized using the following descriptions:

2.2.1 Isolated Equilibrium Point

An isolated equilibrium point is a point in a space that, under a particular transformation, is mapped back onto itself but has no other points mapping onto it.

2.2.2 Asymptotic Equilibrium Point

An asymptotic equilibrium point that, under a particular transformation, is mapped back onto itself, but has other points onto it. The set of points that either map onto an asymptotic equilibrium point, or map onto a point that eventually maps onto an isolated equilibrium point are referred to as the points **basin of attraction.**

2.2.3 Limit Cycle

A limit cycle is a set of points that map to each other cyclically under successive transformations. A limit cycle can either by isolated or asymptotic in the same way as a point.

2.2.4 Unstable Trajectory

An unstable trajectory is one that neither reaches a limit cycle nor an equilibrium point.

3 EXPERIMENTAL PROCEDURE

3.1 Selection of Initial Sentences

English sentences were chosen at random from already published works from a variety of fields including philosophy, biology, mathematics, children's literature, and classical literature. The sentences were then assessed using the Flesch-Kincaid reading ease calculation available in Microsoft Word 2000. A total of 10 sentences with assessed reading ease scores 30 and 98.2 were used.

3.2 Determination of Trajectories

The sentences were entered into the babelizer program developed as freeware by Jonathon Feinberg and available on the internet at http://MrFeinberg.com/babelizer/. Feedback Iterations were performed on the sentences to a maximum of 12 iterations. The results of the iterations were stored in a text file. Iterations for each initial sentence were performed from English to French, German, and Italian, and back again.

3.3 Quantification of Trajectories

The text files were then entered into the program lexicom.exe developed by the author. This program assesses the difference between each iteration of the sentence and the original to determine the distance of the result of each iteration from the original. The algorithm used by lexicom.exe is described in section 4. The numerical results provided by lexicom.exe were then entered into MatLab for analysis.

4 THE LEXICOM PROGRAM

4.1 Purpose

The lexicom program provides a metric in the space of collections of English words. It is used to determine how different one sentence is from another and plot the trajectories of successive iterations with respect to the initial condition sentence. The lexicom provides a true metric as defined in 2.1.1 ,and it is only with this sort of program that one can make sense of the iterations that result from successive translation.

4.2 Compare Method

The compare method is the fundamental core of the lexicom program. It takes two text strings string1 and string 2 and assesses how different they are. It does by searching for every word in string1 in string2. When it finds an occurrence of a word from string1 in string2, it compares the neighbors or this word recursively until both ends of the sentence are reached. The score for each word and each word neighbors is added up to generate the compare score.

4.3 Lexicom Comparison Process

The steps involved in making a lexicom comparison are described below:

1. Strip all white space, punctuation, and formatting characters from both sentences.
2. Compare the first string to the second and store as c1.
3. Compare the second string to the first and store as c2.
4. Compare the first string to itself and store as c3.
5. Compare the second string to itself and store as c4.
6. Determine Length difference factor as Lerror=sqrt(abs((L1-l2)/(L1+L2)))
7. The lexicom score is now sqrt(Lerror*(c1+c2)/(c3+c4))-1

4.4 Another Available Comparison Technique

There exists a method known as latent Symantec analysis that I originally thought would be useful in this project. It takes sentences and compares them based on a database of word relations to determine if they are similar in meaning or scope. The problem with using this method in this circumstance is that most sentences remain very close to each other in topic, and the resolution of this method of comparison is not great enough to differentiate between successive iterations. For more information and online versions of this comparison technique please see http://lsa.colorado.edu.

5 ANALYSIS

5.1 Identification of the Qualitative Properties of the System

5.1.1 Identification of Isolated Equilibrium Points

There is no simple way to prove or disprove the existence of isolated equilibrium points without performing an exhaustive search on all word combinations to determine that there are some to and from which no other points are mapped. During the experiments, certain words were found to be stable and never mapped to, and these are possible candidates for isolated equilibrium points. Such words were often proper names and technical terms, and it is suspected that some of these are true isolated equilibrium points

5.1.2 Identification of Asymptotic Equilibrium Points

The results of iteration of sentence 1 and 4 to and from all languages resulted in convergence to an asymptotic equilibrium point. Sentence 2 ,to and from French, sentence 3, to and from German, and sentence 5, to and from both Italian and German, also result in convergence to an asymptotic equilibrium point.

5.1.3 Identification of Limit Cycles

Several limit cycles exist in the data sets acquired through these experiments. A period-7 limit cycle is visible in figure 6.2.2 and period-2 limit cycles are visible in 6.2.1, 6.2.3, 6.3.2, and 6.5.2.

5.1.4 Identification of Unstable Trajectories.

A single unstable trajectory was discovered and is visible in 6.3.2 The result of successive iterations moves farther and farther away from the original sentence. This instability is the result of the mapping (skate->ice-skate) which produces longer and longer strings of (ice-ice-ice…).

5.2 Assessment of the Dynamic Responses

Linear systems can be assessed by how quickly they converge to stable points. With nearly-linear systems these responses are exponential and can be characterized by the exponent that best fits their responses (which is analytically the exponent of the response after linearizing about that stable point). In the case of many of the responses documented in this paper, the basin of attraction is much larger than the linearization is near the stable point, and the responses can not be fit accurately by exponentials. In cases such as this the response can described by upper and lower bounding it with exponentials. The next few sections apply some properties of almost-linear systems to the results obtained.

5.2.1 Over-shoot

Almost-Linear systems can experience a phenomenon known as overshoot in their trajectories. This phenomenon is visible in 6.1.1, 6.3.3 and 6.5.1. The percentage overshoot for these three results in available in the following table:

Sentence	Overshoot %
1-Italian	30
3-German	100
5-Italian	7.7

5.2.2 Convergence Exponent

The almost linear systems that experience no overshoot can be described directly by curve fitting an exponential to them. Below is a summary of the exponential that best curve fit the response for each of these cases:

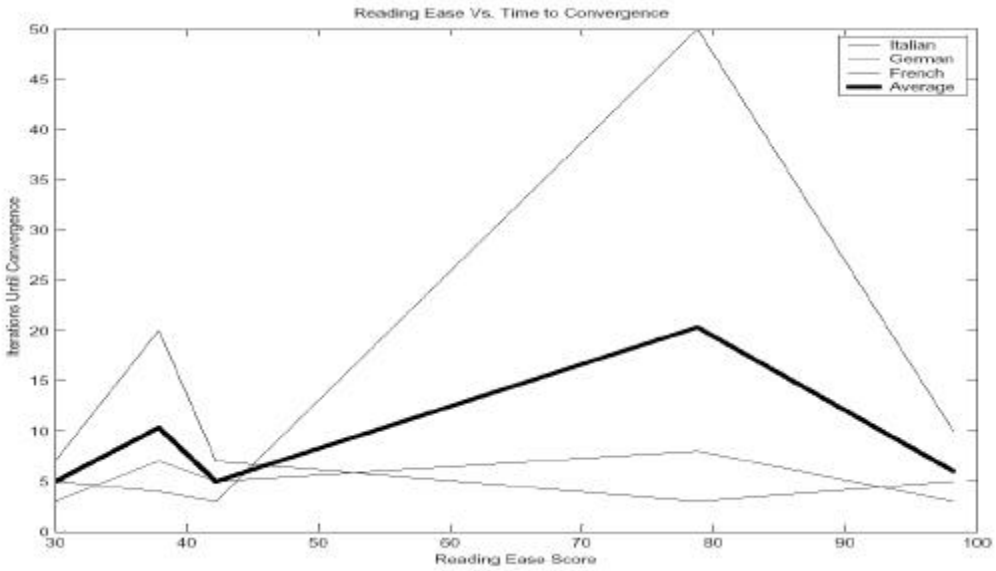
Sentence	Exponent
1-French	2.5
4-French	1.5
4-German	1.25

5.3 Comparison Based on Reading Ease

5.3.1 Reading Ease vs. Convergence

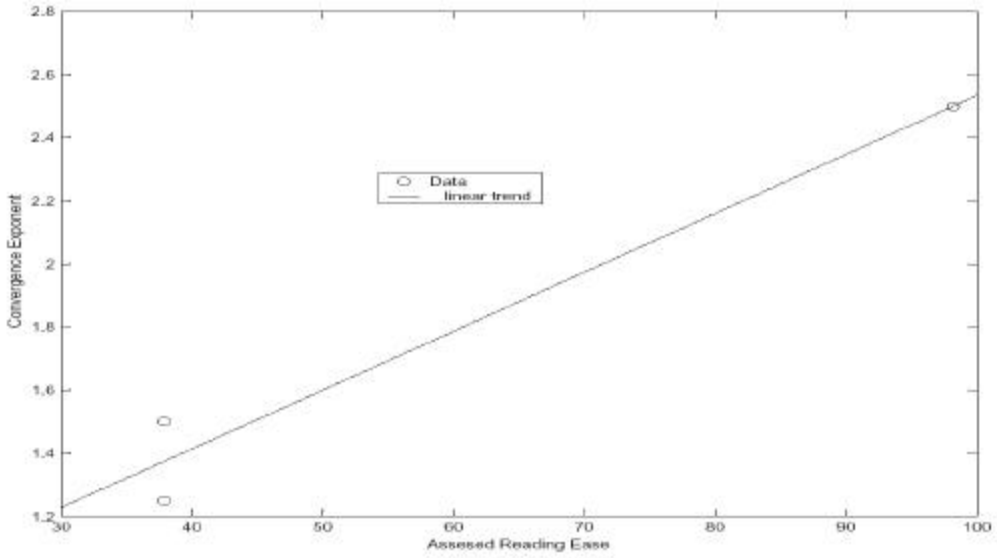
5.3.1.1 Reading Ease vs. Iterations to Convergence

The following plot compares the assessed reading ease of sentences to the time taken for those sentences to converge in different languages. From this plot it is evident that there is no general trend in this way according to the data analyzed in this paper.



5.3.1.2 Reading Ease vs. Convergence Exponent

The following plot shows the relationship between reading ease and the calculated convergence exponent. A distinct trend is apparent for the cases where the region of convergence is nearly-linear and no overshoot exists.



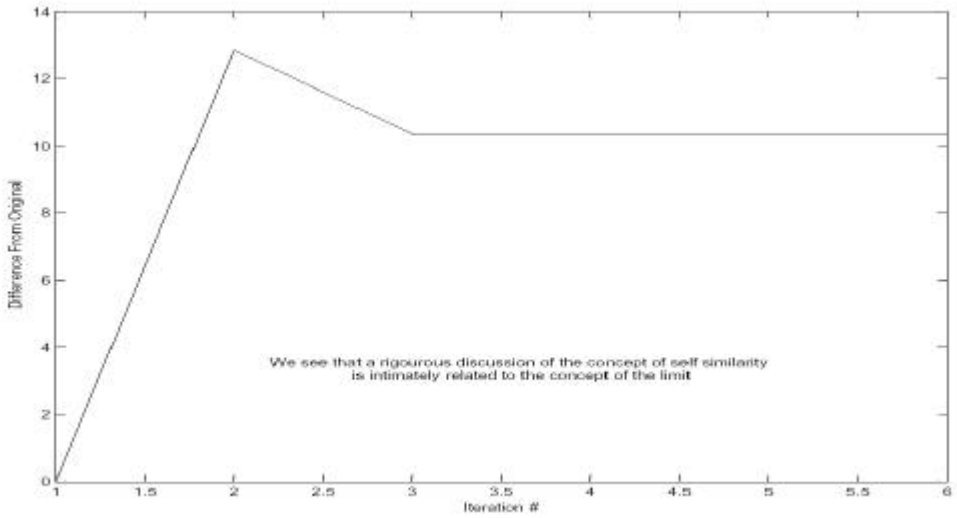
6 SELECTED RESULTS

6.1 Sentence 1

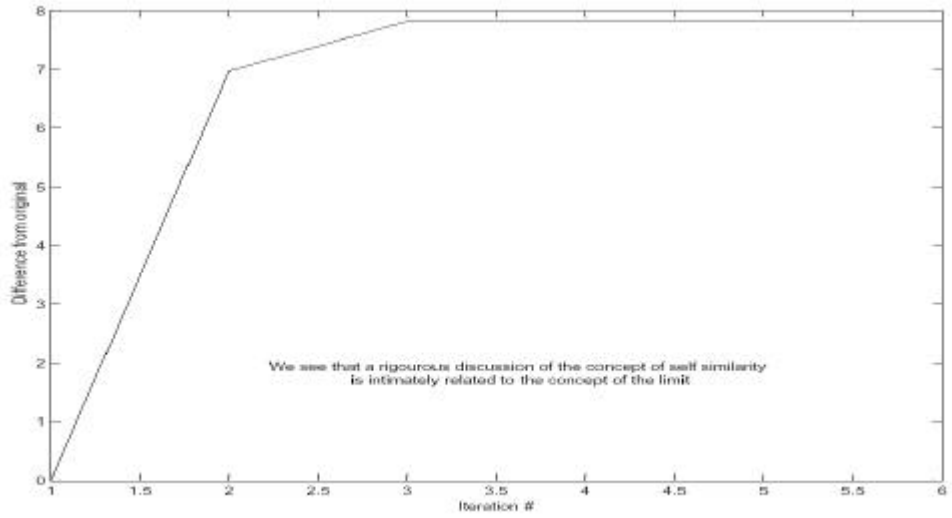
We see that a rigorous discussion of the concept of self similarity is intimately related to the concept of the limit.

Reading Ease: 30

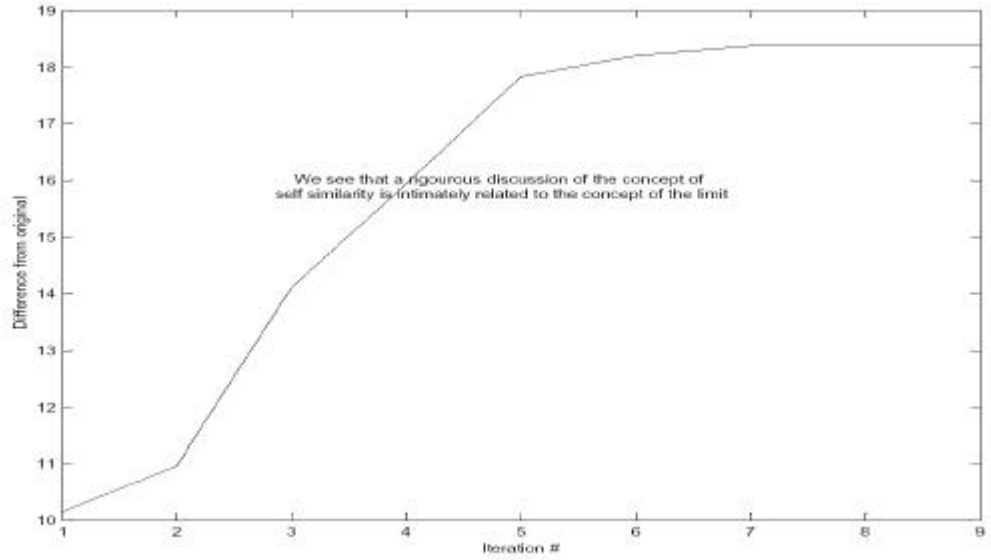
6.1.1 Italian



6.1.2 French



6.1.3 German

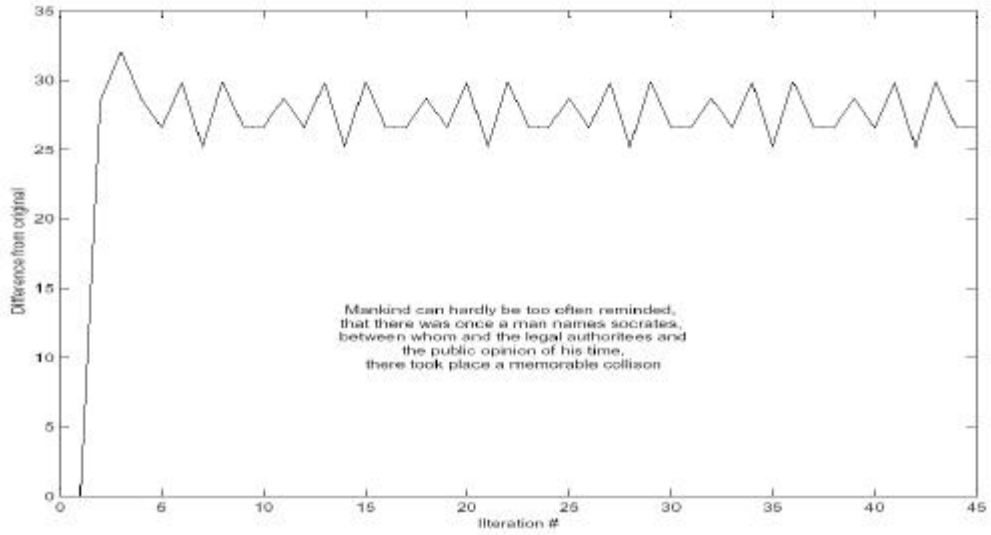

We see that a rigourous discussion of the concept of
self similarity is intimately related to the concept of the limit
Difference from original
Iteration #
19
18
17
16
15
14
13
12
11
10
1
2
3
4
5
6
7
8
9

6.2 Sentence 2

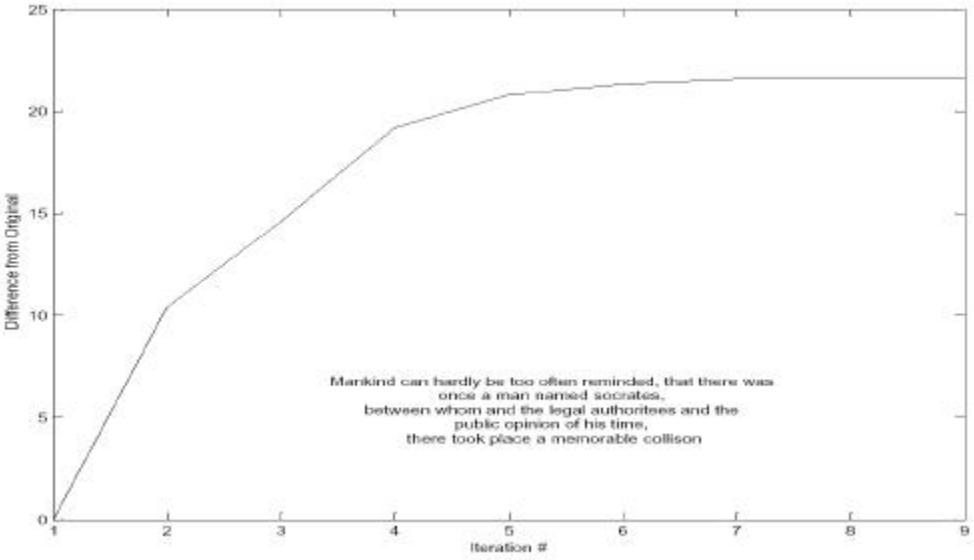
Mankind can hardly be too often reminded, that there was once a man names socrates, between whom and the legal authoritees and the public opinion of his time, there took place a memorable collison

Reading Ease: 37.9

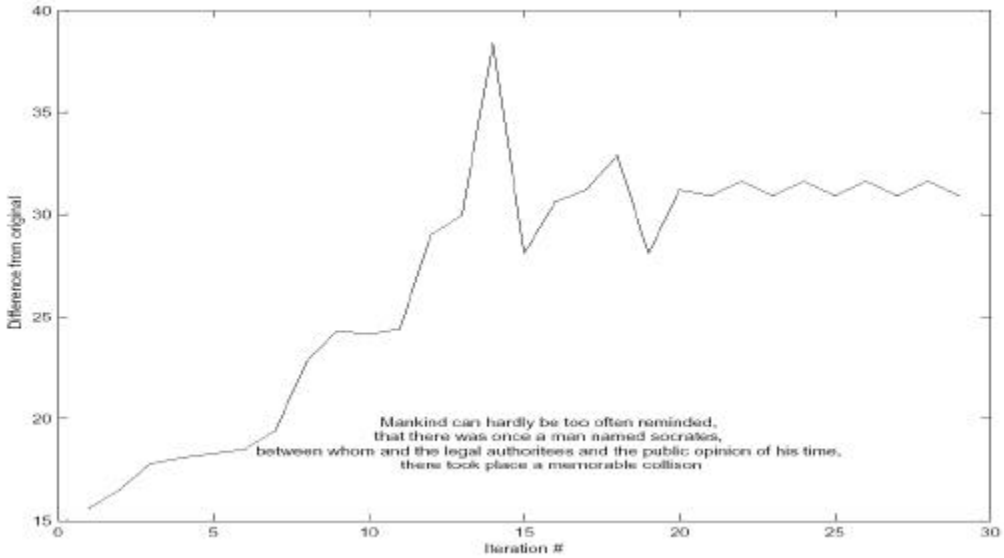
6.2.1 Italian



6.2.2 French



6.2.3 German

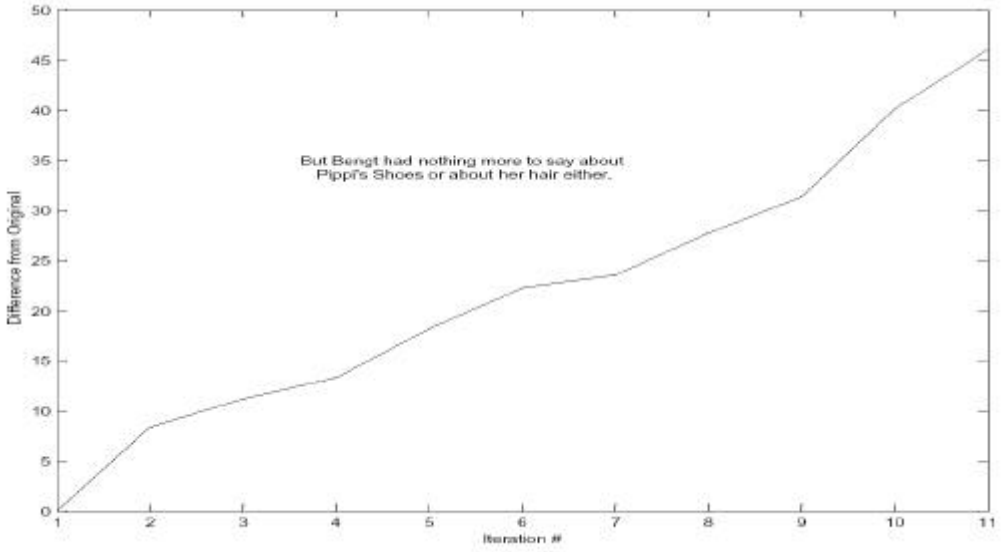

Difference from original
Iteration #
Mankind can hardly be too often reminded,
that there was once a man named socrates,
between whom and the legal authoritees and the public opinion of his time,
there took place a memorable collision

6.3 Sentence 3

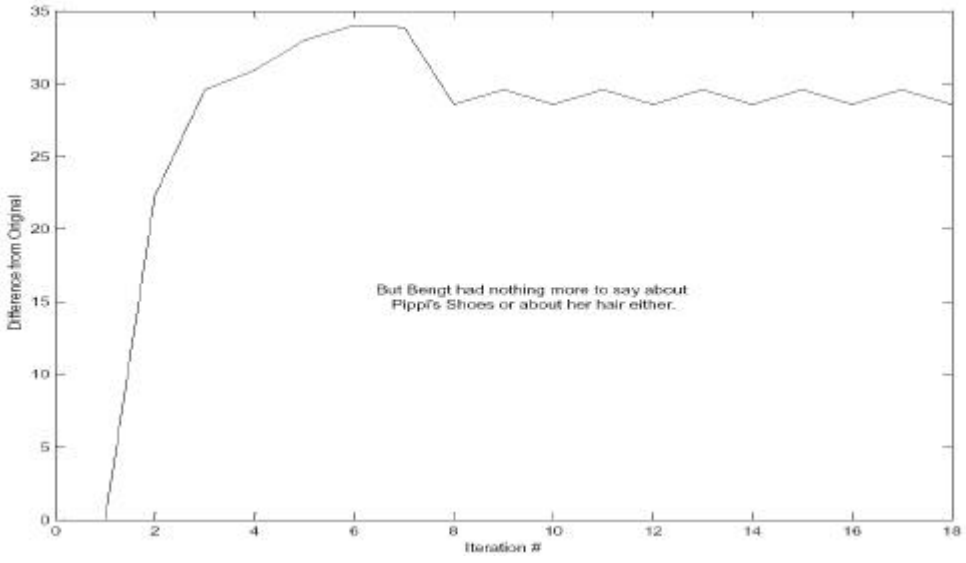
But Bengt had nothing more to say about Pippi's Shoes or about her hair either.

Reading Ease: 78.8

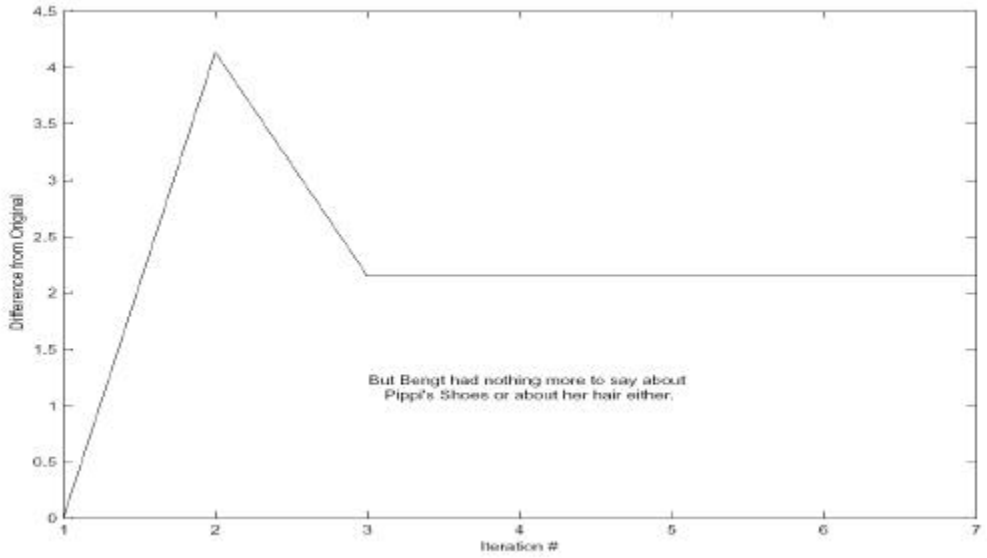
6.3.1 Italian



6.3.2 French



6.3.3 German

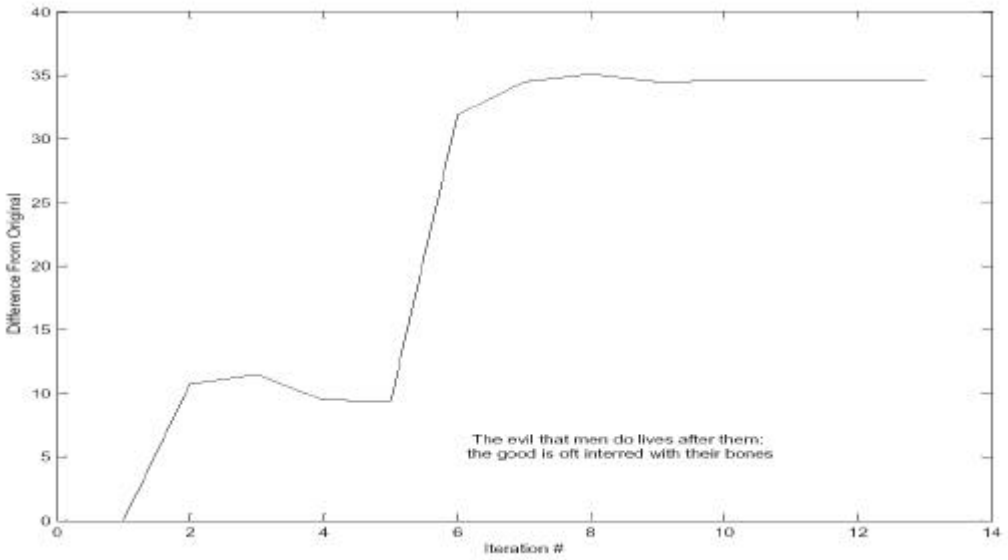


6.4 Sentence 4

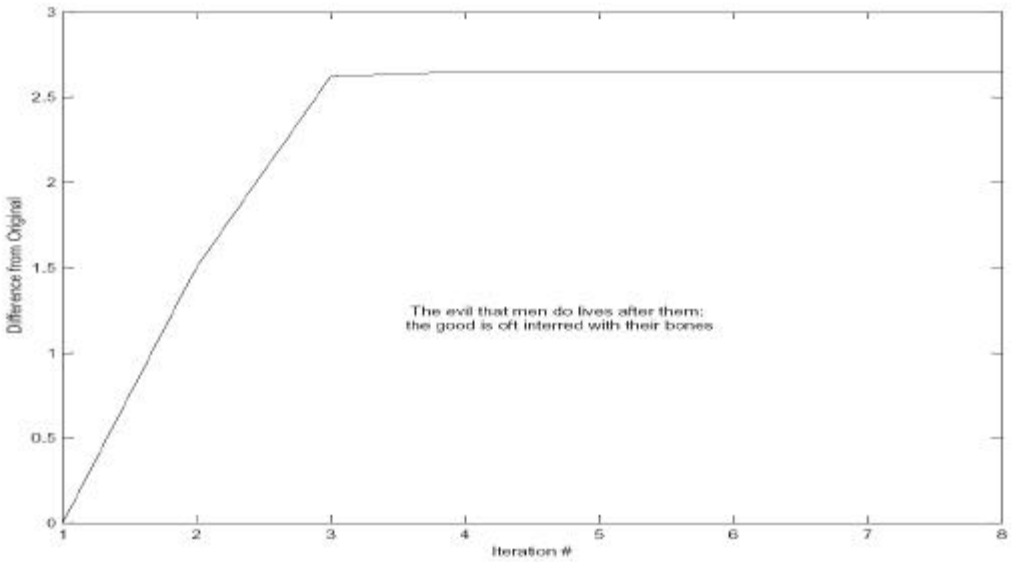
The evil that men do lives after them; the good is oft interred with their bones.

Reading Ease: 98.2

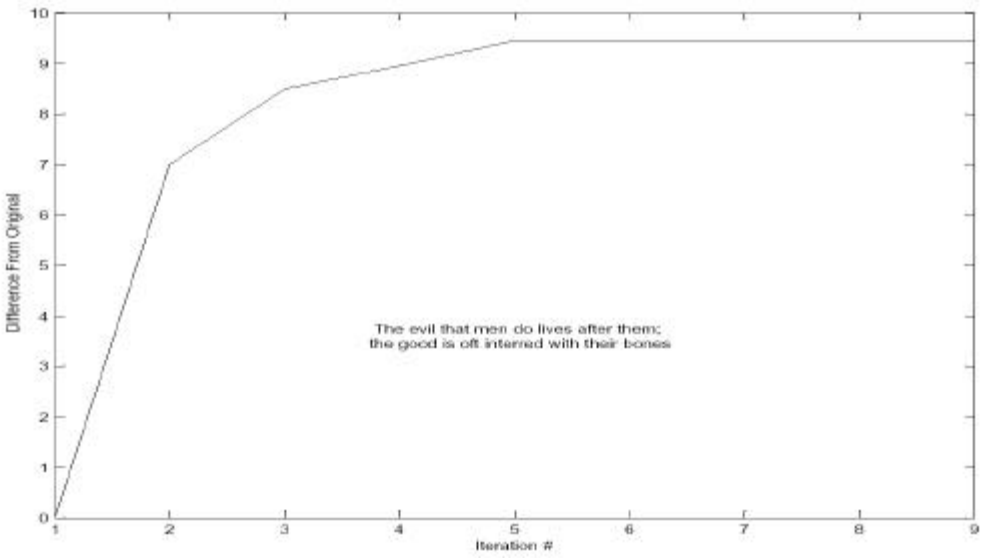
6.4.1 Italian



6.4.2 French



6.4.3 German

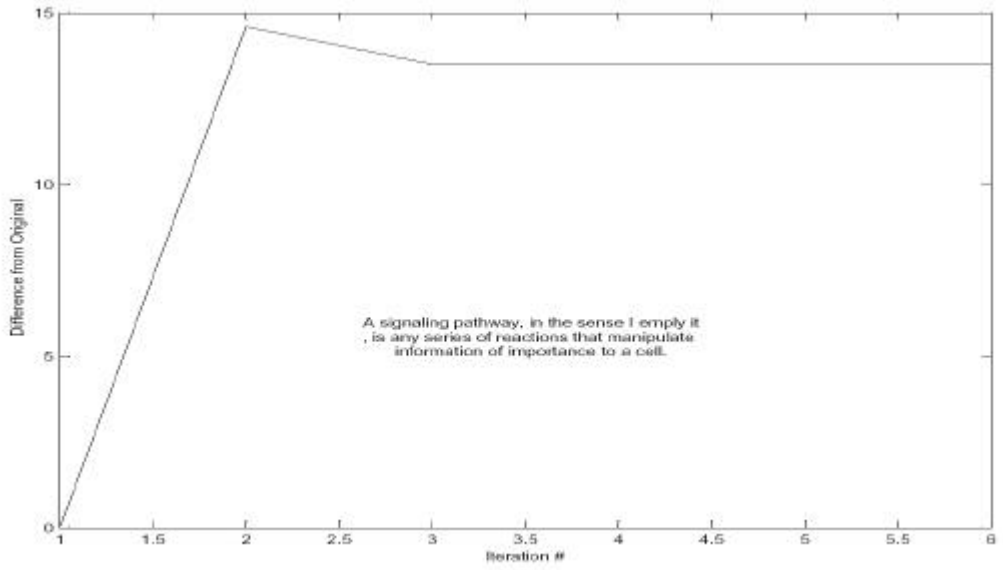


6.5 Sentence 5

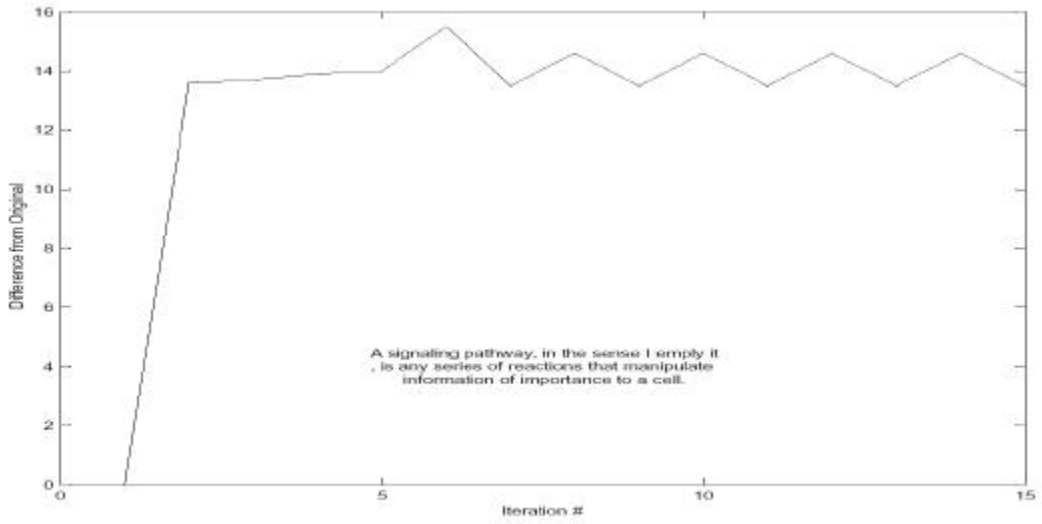
A signaling pathway, in the sense I employ it, is any series of reactions that manipulate information of importance to a cell.

Reading Ease: 42.2

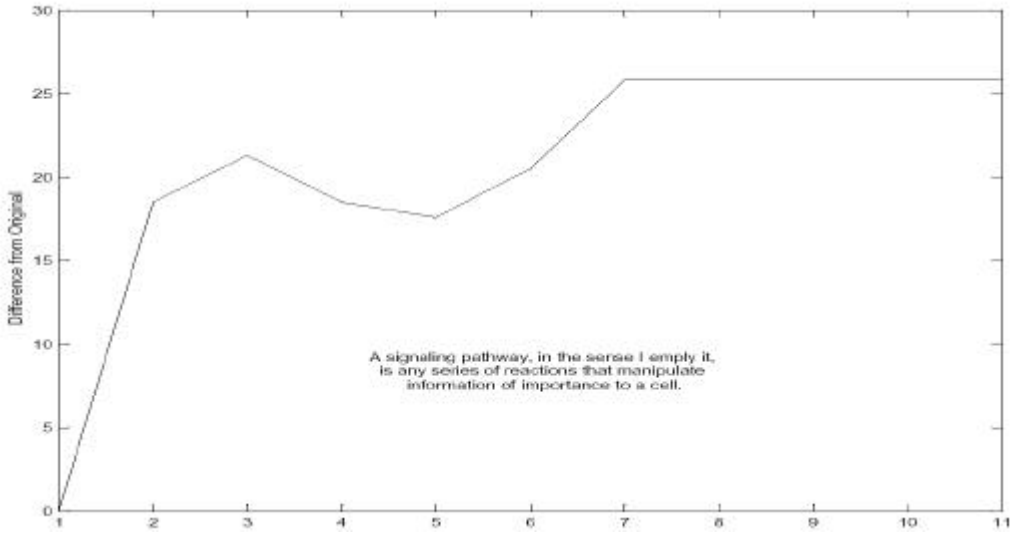
6.5.1 Italian



6.5.2 French



6.5.3 German


Difference from Original
A signaling pathway, in the sense I emply it, is any series of reactions that manipulate information of importance to a cell.
0
5
10
15
20
25
30
1
2
3
4
5
6
7
8
9
10
11

7 FURTHER WORK

7.1 Estimation of the Region of Convergence

Further work can be done to try to estimate the region of convergence of different stable points. This should provide more insight into the topology of the space in question and allow the trajectories of points converging to a single stable point to be compared.

7.2 Assessing the Effect of Other Sentence Properties

The only sentence property quantitatively investigated in this paper was reading ease. Other properties may be more predictive of stability results. Comparing sentences by very tense, number of adjectives and other measures of complexity are only some of the many possibilities.

7.3 A Better Metric

The lexicom program works surprisingly well, but it does have many limitations. The primary one is that it completely ignores meaning in the sentences presented to it. This makes it very difficult for it to assess the progress of individual words that are iterated, or estimate and predict regions of convergence. Perhaps a combination of the current Lexicom program and the techniques available from latent semantic analysis would yield better results.

7.4 Multiple Language Translations

More insight may be obtained by iterating sentences from English to another language, from that language to another and then back to English. This arrangement might aggravate any instabilities in the system and allow them to be more easily identified and classified.

8 CONCLUSION

Although the research presented in this report is very preliminary, it is apparent that iterations of machine translation software do constitute a dynamical system that can be analyzed using relevant techniques. The identification and classification of subsets of the space of English sentences that are stable, unstable, or a member of a stable limit cycle, provides a starting point from which more accurate descriptions of the properties of these subsets can be assessed in greater detail.

Most sentences converge very quickly to stable points or limit cycles. There seems to be no correlation between reading ease and time to convergence, although some relationship between reading ease and convergence exponent is seen. The only unstable trajectory discovered was a result of an initial condition found in children's literature.

The relatively short convergence seen in this paper is encouraging as it means that the space of English sentences is dense with stable points. If this is the case, it should be relatively easy to find translationaly invariant language to suggest to authors during a pre-translation step.

Sentences tend to converge rather quickly to a stable point or cycle, but it seems to be unlikely that a sentence chosen at random will be translationaly invariant, as none of the 10 sentences chosen for this study were such. This is encouraging if one wants to use the stability properties to asses the fidelity of translation software. If most sentences are not stable points, but near stable points, it will not take very long to iterate many different sentences to compare translation software without human intervention.

9 REFERENCES

Arnold, Balkan, Meijer, Humphreys and Louisa Sadler *Machine Translation: an Introductory Guide*, Blackwells-NCC, London:1994

Barnsley, Michael *Fractals Everywhere* Academic Press, INC, San Diego, CA: 1988

Khalil, Hassan *NonLinear Systems* -Prentice Hall, Upper Saddle River, NJ: 1996

Lindgren, Astrid, *Pippi Longstocking,* Puffin Books, New York: 1997

Mill, John Stuart *On Liberty and Other Essays,* Oxford University Press, Oxford:1991

O'Neil, Peter V. *Advanced Engineering Mathematics*, Brooks/Cole Publishing Company. London:1995

Oppenheim, Alan V, Schafer, Ronald W. *Discrete Time Signal Processing,* Prentice Hall, New Jersey: 1999.

Peitgen, Jurgens, Saupe, *Chaos and Fractals,* Springer-Verlag, NY,NY:1992

Rosen, Kenneth *Discrete Mathematics and Its Applications* Mcgraw-Hill: 1999

F. Reike, D. Warland, R. Ruyter van Steveninck, W. Bialek*, Spikes: Exploring the Neural Code***,** MIT Press. Cambridge, Massachusetts: 1998

Shakespeare, William *The Complete Works,* Gramercy Books, New York: 1975