Linguaggio nei numeri e numeri nel linguaggio. Linguistica, Matematica e Cryptonalisi

Thomas Christiansen *Dipartimento Studi Umanistici, Università del Salento, Lecce, Italia*

1 Intoduzione: la scienza della linguistica

Ai tradizionalisti, gli studi di matematica e di linguistica, potrebbero sembrare avere poco in comune, poiché costituiscono due modi di apprendimento molto differenti. Nelle università medioevali le sette arti cosiddette liberali, erano separate in una divisione inferiore o *trivium* (grammatica, retorica, e logica) e una superiore, *quadrivium* (aritmetica, musica, geometria e astronomia). Oggi la matematica è trattata come una scienza. Lo studio del linguaggio è ancora, ma sempre meno, unito ad argomenti come letteratura e traduzione e classificato come una delle arti[1]. Ulteriore confusione emerge dal fatto che il termine linguista viene utilizzato in modi piuttosto differenti: da una parte descrive chi sa parlare bene lingue diverse (cioè un poliglotta o

[1]In maniera analoga, la logica è uscita dagli stretti confini del *trivium* in quattro campi differenti: filosofico, formale, informale e matematico.

1 Introduction: the Science of Linguistics

To the more traditionally minded, the studies of mathematics and linguistics may seem to have little in common, constituting two very different types of learning. In medieval universities, the seven so-called liberal arts were divided into a lower division or *triuvium* (grammar, rhetoric, and logic) and a higher *quadrivium* (arithmetic, music, geometry, and astronomy). Today, maths is treated as a science. Still, but increasingly less so, language is sometimes lumped together with things like literature and translation and classed as one of the arts[1]. Added confusion comes from the fact that the term linguist has quite different uses: on the one hand, it can describe someone who is good at speaking different languages (i.e. a polyglot or plurilingual person); on the other,

[1]Similarly, logic has also moved out of the strict confines of the *triuvium* into four separate fields: philosophical, formal, informal, and mathematical.

una persona plurilingua): dall'altra parte denota lo studioso che analizza i linguaggi e le loro strutture, spesso ad un alto livello di astrazione. Quest'ultimo è uno scienziato, specialmente nella sua incarnazione moderna, poiché osserva, misura, costruisce e verifica ipotesi.

L'incremento di un approccio scientifico alla linguistica è emerso gradualmente a partire dalla metà del ventesimo secolo fino a diventare la corrente principale. Questo è avvenuto rendendosi conto che il linguaggio è un fenomeno naturale, un prodotto della biologia umana (come l'abilità di ridere o di camminare su due gambe), non è un artefatto culturale creato dall'uomo (come la birra, il pane o la poesia). Per questo il linguaggio ha acquisito un significato speciale nei contesti della biologia, delle scienze cognitive e della psicologia poiché oggi è riconosciuto come qualcosa che si è evoluto, come il resto dei nostri corpi e menti, in centinaia, migliaia e possibilmente milioni di anni. Conseguentemente, è qualcosa di molto più complesso e complicato di ogni cosa che possa essere prodotta dall'uomo.

L'argomento che l'abilità di usare il linguaggio (sebbene non si parli di una lingua specifica) sia istintivo è comunemente associata a Chomsky [1], sebbene ci siano indicazioni di questa visione in Darwin (Ref. [2], Chapter III):

> "Come osserva Horne Tooke, uno dei fondatori della nobile scienza della filologia il linguaggio è un'arte come fare la birra o il panificare; ma scrivere sarebbe simile. Non è un vero istinto, poiché ogni lingua deve essere appresa. È tuttavia molto diversa da tutte le arti comuni dato che l'uomo ha una tendenza istintiva a parlare, come vediamo nel balbettare degli infanti; mentre nessun bimbo ha la tendenza a produrre birra, panificare o scrivere." [Sottolineatura nostra].

Prima di rendersi conto che il linguaggio fosse istintivo, la cosidetta `teoria comportamentista` della psicologia considerava che l'acquisizione potesse essere basata puramente su stimoli e risposte. Chomsky, e molti altri dopo lui, tra i quali lo psicologo cognitivo Steven Pinker [3], controbatterono che questi meccanismi non riuscivano

it denotes the type of scholar who analyses languages and their structures, often at highly abstract levels. The latter is a scientist, especially in their modern incarnation, because they observe, measure, make and test hypotheses.

The increasingly scientific approach to linguistics gradually came to dominate the mainstream from the mid-20th century onwards. It arose from the realisation that language is a natural phenomenon, a product of human biology (like the ability to laugh or to walk on two legs), not a cultural artefact created by humans (like beer, bread, or poetry). As such, language is something that has acquired special significance in the contexts of biology, cognitive science and psychology because it is now recognised as something that has evolved, like the rest of our bodies and minds, over hundreds of thousands, possibly millions, of years. Consequently, it is something far more complex and complicated that anything human-made could ever be.

The argument that the ability to use language (though not to speak a specific language) is instinctive is most readily associated with Chomsky [1] although there are hints of this view in Darwin (Ref. [2], Chapter III):

> "As Horne Tooke, one of the founders of the noble science of philology, observes, language is an art, like brewing or baking; but writing would have been a better simile. It certainly is not a true instinct, for every language has to be learnt. It differs, however, widely from all ordinary arts, for *man has an instinctive tendency to speak*, as we see in the babble of our young children; whilst no child has an instinctive tendency to brew, bake, or write." [Emphasis ours]

Prior to the realisation that language was instinctive, the so-called behaviourist theory of psychology had held that acquisition could be based purely on stimulus and response. Chomsky, and many others since, including the cognitive psychologist, Steven Pinker [3], counter that such mechanisms cannot account for the

a tener conto della velocità con la quale ogni bambino acquisisce il linguaggio o la complessità della strutture sintattiche che produce, anche ad una giovane età, alcune di queste sono generate autonomamente da regole astratte e non sono prodotte dall'imitazione di quello che sente attorno a sé. Secondo Chomsky [4], i bambini nascono con una grammatica universale che si può descrivere come `intrinsecamente connessa` nel cervello. Questo è evoluto come meccanismo con cui l'infante può analizzare un numero limitato di *input* di una lingua specifica e convertirlo in un modello mentale funzionante del linguaggio stesso. Questo modello può essere usato per generare strutture permettendo, in effetti, al bambino di cominciare a parlare la lingua in questione[2]. I linguaggi non sono quindi scaricati o appresi. Invece, sono acquisiti per contatto con *input* adeguati, e ricostruiti attraverso la mappatura degli schemi e delle forme incontrati nella cornice fornita dalle grammatica universale che è innata.

Alcuni dei primi studi di linguistica nell'era moderna erano focalizzati sull'origine dei linguaggi e la connessione tra loro: per esempio, nei secoli diciottesimo e diciannovesimo, l'osservazione di William Jones che le lingue classiche dell'antica Grecia e Roma sembravano condividere delle strutture simili con lingue come il Farsi ed il Sanscrito, e la conclusione che che avevano radici comuni; oppure la legge di Jakob Grimm riguardante i cambiamenti sistematici del suono nelle consonanti occlusive / plosive, a cui furono sottoposte quando evolverono nel tardo Proto-Germanico. Nel ventesimo secolo, i linguisti (ad esempio Ferdinand de Saussure, Otto Jespersen, Leonard Bloomfield, Noam Chomsky, Michael Halliday) rivolsero la loro attenzione alla struttura del linguaggio (in astratto) studiando cose come il meccanismo che permette ai concetti di essere condensati in un messaggio e quindi codificati in suoni, parole o strutture sintattiche. In questo modo, l'analisi di tutti i livelli del linguaggio (ad esempio fonologia, morfologia, lessico, grammatica e sintassi) fu vista come qualcosa

[2]In Inglese, con una certa confusione, lo stesso termine *language* è usato sia per le regole e i princìpi che governano il linguaggio sia per indicare la produzione degli elementi e delle strutture linguistiche: quello che de Saussure [5], in francese, chiamava *langue* e *parole* e che Chomsky [6] ha definito come *competence* e *performance*.

speed at which any child acquires language or the complexity of the syntactic structures that they produce even at a young age, some of which they can be seen to generate for themselves out of abstract rules and not merely through imitation of what they hear around them. According to Chomsky [4], children are born with a universal grammar which one could describe as "hard-wired" into the brain. This has evolved to serve as a mechanism whereby infants can analyse a limited amount of input from a specific language and convert it into a functional mental model of that same language. This model can then be used to generate structures, in effect allowing the child to start "speaking" the language in question[2]. Languages therefore are not so much "handed down" or learnt. Rather, they are acquired through contact with adequate input, and reconstructed through the mapping of the patterns and forms encountered onto the framework provided by the innate universal grammar.

Some of the earliest treatments of linguistics in the modern era were focussed on the origin of languages and the connections between them: for example, in the eighteenth and nineteenth centuries, William Jones's observation that the classical Languages of ancient Greece and Rome seemed to share common features with languages such as Farsi and Sanskrit, and his conclusion that they must share common roots; or Jakob Grimm's laws regarding the systematic sound changes which the Proto-Indo-European stop consonants underwent as they developed in later Proto-Germanic. In the twentieth century, linguists (e.g. Ferdinand de Saussure, Otto Jespersen, Leonard Bloomfield, Noam Chomsky, Michael Halliday) turned their attention to the structure of language (in the abstract), looking at such things as the mechanisms that allow concepts to be put together into messages and then encoded as sounds, words or syntactic structures. In this way, analysis of all levels of language (e.g. phonology, morphology, lexis, grammar, and syntax) came to be seen as something descriptive.

[2]In English, confusingly, the same term, *language*, is used for both the rules and principles governing language, and for the actual production of linguistic elements and structures: what de Saussure [5], in French, called *langue* and *parole* or what Chomsky [6] named *competence* and *performance*.

di descrittivo. Tradizionalmente, le regole del linguaggio sono state considerate come prescritte: non emergenti naturalmente attraverso l'evoluzione del sistema, ma determinate da una convenzione tra gli utilizzatori delle lingue e le generazioni che li hanno preceduti. Per questo motivo i grammatici tradizionalisti e quelli che hanno ancora propensioni pedagogiche, possono scartare con leggerezza quelle cose che non riescono a spiegare come fossero `eccezioni`, senza rendersi conto che stanno falsificando le loro stesse ipotesi.

Oggi, la linguistica è vista come uno studio scientifico di un fenomeno naturale: essenzialmente un'indagine del comportamento, allo stesso modo in cui uno zoologo può osservare un animale oggettivamente e spassionatamente senza ricorrere ad alcuna idea preconcetta di giusto o sbagliato, cioè, come la creatura in questione dovrebbe, o non dovrebbe, comportarsi. Secondo il pluritalentuoso Chomsky[3], che nella prima parte della sua carriera fu all'avanguardia collaborando in vari articoli di linguistica matematica, l'importanza del linguaggio risiede nel fatto che costituisce una "finestra della mente"[4]. Il linguaggio è così centrale per la condizione umana e per la società che non dovrebbe essere una sorpresa che oggi esista una miriade di filoni della linguistica che vanno da modelli astratti fino a studi del linguaggio applicato ad aree specifiche[5]. Ci sono anche numerose e molto diverse maniere di descrivere il linguaggio e le sue strutture, ad esempio, semplicemente collegando le strutture sintattiche: grammatica dei casi, prospettiva funzionale delle frasi, grammatica funzionale al lessico, teoria ottimale, grammatica sistemica, e, in ultimo, ma non meno importante, la grammatica

[3]Anche se è spesso identificato come "padre della linguistica moderna" è anche noto come filosofo, scienziato cognitivo, storico e attivista politico.

[4]Chomsky ritiene che la linguistica sia una branca della psicologia "Sono primariamente interessato alla possibilità di imparare qualche cosa, dallo studio del linguaggio, che porterà a chiarire alcune proprietà inerenti la mente umana. " ([7], p. 90).

[5]Esempi del secondo questo ricadrebbero sotto la generica definizione di lunguistica (ad esempio, lo studio dell'insegnamento del linguaggio); esempi del primo includerebbero approcci molto diversi come: linguistica clinica, linguistica computazionale, linguistica dello sviluppo, ecolinguistica, linguistica evolutiva, linguistica forense, linguistica storica, neurolinguistica, sociolinguistica.

Traditionally, the "rules" of language had been seen as prescribed: not naturally occurring through the evolution of a system, but determined by convention among the users of the languages and the generations that had preceded them - and this is why traditional grammarians, and still some of those today of a pedagogical bent, can glibly dismiss those things that they cannot explain as "exceptions" without realising that they were thereby falsifying their own hypotheses.

Today, linguistics is seen as a scientific study of a naturally occurring phenomenon: essentially an enquiry into behaviour, in the same way that a zoologist might observe an animal objectively and dispassionately without recourse to any preconceived idea of correct and incorrect, that is, how the creature in question should or should not behave. According to the multitalented Chomsky[3], who in his early career was trailblazing in that he collaborated on various papers specifically on mathematical linguistics, the importance of language stems from the fact that it constitutes a "window on the mind"[4]. Language is so central to the human condition and society that it should come as no surprise that today there exist myriad strands of linguistics ranging from abstract models to the study of language applied to specific areas [5]. They are also numerous and very diverse ways of describing language and its structures, for example, just relating to syntactic structures: case grammar; functional sentence perspective; lexical functional grammar; optimality theory; systemic grammar, and, last but not least, Chomskyan transformational-generative grammar.

[3]As well as being often referred to as "the father of modern linguistics", he is also a noted philosopher, cognitive scientist, historian, and political activist.

[4]Chomsky argues that linguistics is a branch of psychology: "I am primarily intrigued by the possibility of learning something, from the study of language, that will bring to light inherent properties of the human mind" ([7], p. 90).

[5]Examples of the latter, would fall under the general banner of applied linguistics (e.g. the study of language teaching); examples of the former would include such diverse approaches as: clinical linguistics; computational linguistics; developmental linguistics; ecolinguistics; evolutionary linguistics; forensic linguistics; historical linguistics; neurolinguistics; sociolinguistics.

chomskiana generativo-trasformazionale.

In questo articolo, fornirò degli esempi di come alcuni aspetti della linguistica si basino su principi matematici. La matematica può essere vista come rilevante in ogni area e ad ogni livello della moderna analisi linguistica. Di conseguenza la sfida, in un breve articolo come questo risiede non tanto nel trovare esempi ma piuttosto nell'evitare di rimanere impantanati dall'evidenza. Per questa ragione stutturerò la presentazione attorno ad una specifica area che illustra, anche a coloro che potrebbero nutrire scarso interesse in problemi di linguistica astratta, l'applicazione pratica della matematica alla linguistica, specificamente la `criptoanalisi` o la decodifica dei codici[6]. Questa è un'area che mostra ampiamente l'alto livello che può raggiungere lo sforzo dell'uomo quando studiosi con diverse conoscenze, tra i quali matematici e linguisti, condividono le loro risolse molto differenti e lavorano insieme per uno scopo comune.

2 Criptoanalisi: decodifica dei codici

Una delle aree più note in cui i percorsi dei matematici e dei linguisti si sono incrociati è nel mastodontico sforzo fatto nel periodo della Seconda Guerra Mondiale dagli Alleati per decifrare le comunicazioni radio criptate delle potenze dell'Asse all'interno dell'area che i militari chiamano *Signal Intelligence* (SIGINT). Questo fu sviluppato non solo dalla britannica `Government Code and Cypher School (Stazione X)`, ma anche: dal prebellico polacco *Biuro Szyfrów* (`Ufficio codici`) (fino al 1939); dal francese *Deuxième Bureau* (fino al 1940); o, negli USA, dal `Army Signal Intelligence Service` del OP-20-G della marina militare. In queste strutture, esperti in vari campi, anche lingue (specialmente quelle usate dal nemico) erano affiancati da matematici. Per ironia del destino questo era in aperta contraddizione con Godfrey Hardy, il pacifista e matematico, che nel 1940 dichiarò pubblicamente che "la vera matematica non ha effetti sulla guerra." [8].

Lavorando insieme come squadre interdisciplinari furono capaci di decifrare codici che erano

[6]Da non confondere con la `criptografia` che è la scienza di produrre codici.

In this article, we will provide some examples of how aspects of linguistics rely on mathematical principles. Mathematics can be seen as relevant to every area and every level of linguistic analysis in modern times. Consequently, the challenge in a brief article such as this lies not in finding examples but rather in avoiding being swamped by the evidence. For this reason, we will structure our discussion around one specific area which illustrates, even to those who may otherwise show little interest in abstract linguistic matters, the practical applications of mathematics to linguistics, namely `cryptanalysis` [6] or codebreaking. This is an area which amply shows the heights of human endeavour that can be achieved when scholars of different backgrounds, among whom mathematicians and linguists, pool their very different resources and work towards a common goal.

2 Cryptanalysis: Codebreaking

One of the best known areas where the paths of mathematicians and linguists have crossed is in the mammoth efforts made in the general period of the Second World War by the Allies to decipher the encrypted radio communications of the Axis Powers within the area of what the military call `Signals Intelligence` (SIGINT). This was carried out not just in Britain's `Government Code and Cypher [sic] School` ("Station X") at Bletchley Park, where Alan Turing worked, but also: the pre-war Polish *Biuro Szyfrów* "Cipher Bureau" (until 1939); the French *Deuxième Bureau* (until 1940); or in the USA, the Navy "OP-20-G", or `Army Signal Intelligence Service`. In such establishments, experts in various fields, including languages (especially those used by the enemy) were paired with mathematicians. In an irony of destiny, this was in direct contradiction of Godfrey Hardy, the pacifist and mathematician, who in 1940 famously declared that "real mathematics has no effect on war." [8].

Working together, such interdisciplinary teams were able to crack codes that were thought by their creators and users to be all but indecipher-

[6]Not to be confused with `cryptography` - code making.

stati pensati dai loro creatori come indecifrabili. Ci riuscirono identificando delle strutture sempre più complesse nei sempre più sofisticati codici generati dalle macchine usate dalle potenze dell'Asse. Per fare questo, e alla velocità richiesta dalle operazioni di guerra, dovettero progettare e costruire (nessuno di questi era un compito facile) della macchine di decifrazione come le `bombe` (uno sviluppo delle polacche *bomba kryptologiczna*) a Bletchley Park e anche il primo computer elettronico programmabile: `Colossus` (si veda la Sez. 3.2).

È possibile decifrare i codici di linguaggio proprio perché i messaggi, il modo in cui il linguaggio si manifesta, contengono, inevitabilmente, delle caratteristiche che sono ripetute e che possono essere contate e osservate a diversi livelli di frequenza in relazione tra loro. Nelle prossime due sezioni introdurrò le regole fondamentali del linguaggio per cifrare (criptografia) e decifrare (criptoanalisi) come preliminari necessari per una discussione più dettagliata della criptoanalisi dei codici altamente sofisticati usati nella Seconda Guerra Mondiale (Sez. 3.0).

2.1 Criptografia

I primi codici conosciuti risalgono ai tempi antichi, ed erano estremamente elementari rispetto agli standard moderni. Per questo motivo lo sforzo di nascondere il messaggio era paragonabile a quello per criptarlo. Nella moderna era delle telecomunicazioni, nascondere un messaggio è sempre più difficile. Prima della Seconda Guerra Mondiale, gli inglesi avevano investito fortemente nella tecnologia per cogliere, ricevere e leggere segnali radio, ed erano capaci di sentire le comunicazione dell'Asse molto più lontano di quanto i Tedeschi e gli Italiani sospettassero, anche nel fronte orientale, che i Tedeschi immaginavano fosse al sicuro al di là del loro raggio d'azione.

In realtà, sfruttando il fatto che i segnali radio viaggiano su distanze estremamente lunghe, durante la Guerra Fredda le cosidette `numbers stations` erano allestite in modo da trasmettere sequenze di numeri apparentemente casuali letti ad alta voce (sopratutto da voci sintetiche). Si pensa che questo metodo sia stato utilizzato da varie agenzie di spionaggio [8] per comunicare con agenti sul campo (a cui veniva indicato un

able. They did this by identifying increasingly complex patterns in the ever more sophisticated machine-generated codes used by the Axis powers. To do this, and at the speed that wartime operations required, they had to design and build (neither easy tasks) decryption machines, such as, at Bletchley Park, the `bombes` (a development of the Polish *bomba kryptologiczna*) as well as the world's first programmable electronic computer: `Colossus` (see § 3.2).

Decryption of coded language is possible precisely because messages, the way that language manifests itself, inevitably consist of repeated features that may be counted and observed in different levels of frequency in relation to each other. In the next two sections, we will introduce the basics of language encryption (cryptography), and decryption (cryptanalysis) as a necessary preliminary for a more detailed discussion of the cryptanalysis of the highly sophisticated ciphers employed in WW2 in § 3.0.

2.1 Cryptography

The earliest known codes go back to ancient times, and they were extremely simple by modern standard. For this reason, as much effort went into hiding the message, inventing secret compartments etc., as into the actual encryption. In the modern era with the use of telecommunications, hiding a message is increasingly difficult. Before WW2, the British had invested heavily in technology to pick up, receive and read radio signals, which meant that they were able to listen in to Axis communications from much further away than the Germans or Italians suspected, even from the Eastern Front, which the Germans had thought were safely out of their range.

Indeed, exploiting the fact that radio signals could travel extremely long distances, during the Cold War so-called `numbers stations`, were set up openly transmitting apparently random series of numbers read aloud (mostly by synthetic voices). This method is widely thought to have been used by some agencies [8] to communicate with agents in the field (who would be given specific times at which to listen for messages),

istante preciso in cui ascoltare i messaggi), e forse qualcuna di queste stazioni trasmette ancora. La grande maggioranze dei numeri trasmessi sono casuali, ma altri, presumibilmente, formano un messaggio cifrato. Gli operatori SIGINT, si presume, sono obbligati a cercare di decifrare tutto, senza sapere quale sia la parte interessante[7].

Al livello più semplice, il criptaggio implica un metodo detto `shift` o `cifrario di Cesare` dal nome di uno dei primi che lo ha adottato: lo stesso Giulio Cesare. Il metodo implica di rimpiazzare una lettera del testo base (quello non cifrato) con un altro carattere o simbolo nel testo cifrato. Ad esempio, A diventa D, B diventa E, C diventa F, e così via. "*We attack at dawn*" quindi diventa "zh dwwdfn dw gdzq". Questo semplice metodo fu utilizzato per secoli aumentando la sofisticazione per renderlo più difficile da decifrare. Ad esempio, si può mescolare, o idealmente randomizzare, lo `shift` (A diventa C, B diventa Z, C diventa P.)

Al di là del sistema di base, il `cifrario di Vigenère` fu sviluppato nel XVI secolo. Questo usa una sostituzione polialfabetica e alfabeti con sostituzione multipla. La codifica del testo originale è fatta usando il `quadrato`, o `tavola`, di `Vigenère` dove diverse colonne indicano diversi `shift` corrispondenti a diverse lettere dell'alfabeto. Tipicamente una parola chiave (ad esempio `water`) indica in quale ordine i diversi `shift` devono essere applicati (ad esempio, nella prima colonna A=W, in quella successiva A=A, così via, e nella terza A=T ecc.). Spesso la parola chiave è ripetuta in tutte le colonne (W A T E R W A T E R W A T E R). In caso di parole chiave più corte, questo è poco sicuro poiché la ripetizione di ogni chiave è frequente, creando sequenze identificabili a criptoanalisti allenati. In alternativa, il resto dell'alfabeto può essere presentato dopo la parola chiave, ad esempio W A T E R B C D F G H I J K L M N O P Q S U V X Y Z. In quest'ultimo caso, la ripetizione è garantita solo dopo 26 lettere, ma questo, essendo la lunghezza esatta dell'alfabeto, può essere un intervallo piuttosto ovvio per il criptoanalista.

[7] Il ministro Cecoslovacco degli interni ha dichiarato che la Cecoslovacchia usava questo sistema per lo spionaggio. Bletchley Park e la base britannica a Cipro sono state entrambe citate come sorgenti di ben note `number station`.

and perhaps still is as some of these stations remain on air. Most of the numbers transmitted are purely random, but others presumably form encrypted messages. SIGINT operators, one presumes, are obliged to try to decipher everything without knowing which parts are of interest[7].

At the basic level, encryption comprises the method known as the `shift` or `Caesar cipher`, after one of its early adopters: Julius Caesar himself. This involves replacing a letter in the plaintext (unencrypted message) with another character or symbol in the ciphertext. For example, A becomes D, B becomes E, C becomes F, and so on. "We attack at dawn" thus becomes "zh dwwdfn dw gdzq". This basic method was used for centuries with increasing sophistications to make it more difficult to crack. For example, one can use scrambled, or ideally randomised, shifts (A becomes C, B becomes Z, C becomes P).

Out of this basic system, the `Vigenère cipher` was developed in the 16th century. This uses polyalphabetic substitution and multiple substitution alphabets. The encryption of the original text is done using a `Vigenère square` or `table` - where different columns give different shifts corresponding to different letters of the alphabet. Typically, a key word (e.g. water) indicates in which order the various shifts are to be applied (e.g. in the first column A = W; in the next, A = A and so on; and in the third, A = T etc.). Often the keyword is repeated in all the columns (W A T E R W A T E R W A T E R). In the case of shorter keywords, this is less secure because the repetition of each key is frequent, creating patterns discernible to a trained cryptanalyst. Alternatively, the rest of the alphabet may written out the after keyword, e.g. W A T E R B C D F G H I J K L M N O P Q S U V X Y Z. In the latter case, a repetition only after every 26 letters is guaranteed, but this, being the precise length of the alphabet, may be a rather obvious interval for a cryptanalyst.

[7] The Czech Ministry of the Interior has declared that Czechoslovakia used them for espionage. Bletchley Park and the British base of in Cyprus have both been cited as the source of one well-known numbers station.

Ci sono cifrari poligrafici che coinvolgono diverse combinazioni di lettere (ad esempio, "th", "er", "wh", "en" o "ing" in inglese) o anche suoni (ad esempio il suono /ð/ in *that, those, thing* o /θ/ in *north, thorough, three* che sono sostituiti da caratteri separati. Alcuni codici hanno mescolato diversi metodi per rendere più difficile la decifrazione, ad esempio i `nomenclator`[8] usati nell'alto Medio Evo fino ai primi dell'800 che usavano `shift` e un sistema di simboli con combinazioni di lettere, suoni, parole comuni (ad esempio *he, who, is, have, that*) e nomi propri. Ad esempio il nostro messaggio cifrato "We attack at dawn", "zh dwwdfn dw gdzq" può essere reso più complesso come "£ dw*d €+ gr %", dove £ è il pronome *we*, * indica che la lettera precedente deve essere ripetuta, € è la comune combinazione di lettere *ck*, % sta per *wn* e + è la preposizione *at*.

La conoscenza della lingua codificata[9] permette di indovinare i possibili significati di certi gruppi di caratteri nel testo cifrato. Si può essere guidati da cose semplici come la lunghezza della parola. Ad esempio nel nostro testo cifrato "zh dwwdfn dw gdzq" si può fare una ragionevole ipotesi sul possibile significato di "zh" e "dw" dato che c'è un numero limitato di parole di due lettere in inglese (ad es. *at, ax, in, so, to, we, of*) e alcune sono molto più frequenti di altre (*at* e *ax*). Ogni gruppo di tre lettere in un testo cifrato in inglese ha altissima probabilità di essere *the* oppure *and*, dato che sono molto più frequentemente usate di altre parole di tre lettere (*for, not, you* ecc.)[10]. La lunghezza delle diverse parole può essere mascherata raggruppandole in gruppi, ignorando le spaziature. In questo modo il nostro esempio potrebbe diventare "zhdww dfndw gdzqy" (la y è un carattere casuale aggiunto proprio per completare il gruppo di 5 caratteri)

[8] Così chiamati perché usati come pizzini dal `nomenclator` (l'ufficiale che declamava i nomi egli ospiti onorevoli che arrivavano per funzioni importanti.

[9] A volte, per aumentare la sicurezza, varie lingue poco comuni possono essere usate: ad esempio, le forze armate americane impiegavano *speaker* di lingue poco studiate, come Comanche, Navajo o Basco (Euskara), come telefonisti e operatori radio nella prima e seconda guerra mondiale. Più recentemente l'esercito Britannico ha usato il gallese per comunicazioni non vitali nelle operazioni NATO di pace nella ex-Yugoslavia.

[10] Si veda, ad esempio, Oxford English Corpus (www.sketchengine.eu/oxford-english-corpus/).

There are also polygraphic ciphers that involve also different combinations of letters (for example, "th", "er", "wh", "en" or "ing" in English) or even sounds (e.g. the /ð/ sound in that, those, thing, or the /θ/ in north, thorough, three) being substituted by separate characters. Some codes have also used mixtures of different methods to make deciphering more difficult, such as the `nomenclators`[8] used in the later Middle Ages until the 1800s, which used shifts and a system of symbols for common combinations of letters, sounds, and common words (e.g. the, who, is, have, that) and proper names. For example, our cipher of "We attack at dawn", "zh dwwdfn dw gdzq" could be enhanced as "£ dw*d €+ gr %", where £ stands for the pronoun we, * indicates that the previous letter is doubled, € stands for the common letter combination of ck, % for wn, and + for the preposition at.

Knowledge of the features of the language encoded[9] also allows one to guess at the possible meanings of certain groupings of characters in the ciphertext. One may be guided by such simple things as the length of the word. For example, in the ciphertext "zh dwwdfn dw gdzq" above, one may take an educated guess at the meaning of "zh" and "dw" merely because there are a limited number of two letter words in English (e.g. *at, ax, in, so, to, we, of*) and some are far more frequent than others (cfr. *at* and *ax*). Any three letter word in a ciphertext based on English has a very high probability of being either the or and, so frequently are they used in relation to the other three letter words (*for, not, you* etc.)[10]. The length of different words can be disguised by grouping the symbols in the ciphertext into uniform groupings, ignoring word-spaces. In this way, our example could become "zhdww dfndw gdzqy" (y being a random character added just to

[8] So called because they were often used as crib sheets by nomenclator (the official who called out the names of honoured guests arriving at important functions).

[9] Sometimes to increase security, various unfamiliar languages could be used: for example, the US military employed code talkers, speakers of lesser-studied languages, such as Comanche, Navajo or Basque (Euskara), as telephonists and radio operators in WW1 and WW2. More recently, the British army is said to have sometimes used Welsh for non-vital communications in its NATO peacekeeping operations in Ex-Yugoslavia.

[10] See, for example, Oxford English Corpus (www.sketchengine.eu/oxford-english-corpus/)

o "£dw*d €+ gd %".

2.2 Criptoanalisi

Tutti i metodi di criptaggio citati nella sezione precedente sono stati resi notevolmente meno sicuri con l'applicazione di quello che oggi è denominata `frequency analysis`. Questo metodo risale almeno al IX secolo e ad Al-Kindi nel suo "Manoscritto per decifrare messaggi criptati" [9]. Nel tardo XV secolo, lo stesso approccio ha raggiunto l'Europa, con Cicco Simonetta che ha scritto un manuale che utilizza metodi simili.

La `Frequency analysis` è molto efficace con i classici codici descritti nella sez. 2.1 perché, nelle sequenze del linguaggio scritto, alcune lettere e loro combinazioni appaiono con varia frequenza. Ad esempio, la lettera `e` è più comune delle lettere `x` (che è il motivo perché a `Scrabble` ci sono più tessere di certe lettere comuni, e sono maggiori i valori delle lettere più rare). Inoltre, alcune combinazioni di lettere appaiono più frequentemente di altre (ad esempio i digrammi `an`, `er`, `on` e `th` sono alcune delle coppie di lettere più comuni in Inglese). Eccetto in alcuni nomi propri stranieri o parole derivate da altre lingue, alcune combinazioni non appaiono affatto, ad esempio in Inglese: `hj`, `bq`, `xz`. In modo simile alcune lettere sono spesso raddoppiate (ad esempio `ee`, `ff`, `ll`, `oo`, `ss`, `tt`) mentre altre non lo sono mai (ad esempio `aa`, `ii`, `hh`, `jj`, `qq`, `vv`, `xx`). Inoltre, alcune lettere si trovano in precise posizioni, ad esempio, in Inglese, è estremamente raro trovare più di tre consonanti in sequenza; la lettera `y` è più frequentemente trovata alla fine di una parola che all'inizio.; solo una manciata di parole termina con `i`[11].

Il fatto che parole e lettere siano combinate in maniera strutturata e prevedibile è stato sfruttato dal famoso scrittore per l'infanzia Charles Lutwidge Dodgson (noto come Lewis Carrol), autore dei libri di "Alice nel paese delle meraviglie". Non è una coincidenza che Dodgson fosse assunto come insegnante di matematica al `Christchurch Colleague` di Oxford (specializzato nei settori della geometria, logica matematica, algebra li-

[11] La stessa analisi può essere applicata ai suoni: il suono distintivo `ng` / `ŋ` / come in *thing*, non appare mai alla fine di una parola.

complete the last 5-character group), or "£dw*d €+ gd %".

2.2 Cryptanalysis

All the encryption methods cited in the previous section were rendered notably less secure with the application of what is now called `frequency analysis`. This dates back to at least the 9th century and Al-Kindi in his "A Manuscript on Deciphering Cryptographic Messages" [9]. In the later 15th century, the same approach had reached Europe, with Cicco Simonetta also writing a manual employing similar methods.

`Frequency analysis` proves effective with the classical ciphers described in § 2.1 because, in a stretch of written language, certain letters and combinations of letters occur with varying frequencies. For example, the letter `e` is more common than the letter `x` (which is precisely why in a Scrabble set there are more of tiles of certain common letters, and the values of rarer letters are higher). Furthermore, certain combinations of letters occur more frequently than others (e.g. the bigrams or digraphs of `an`, `er`, `on` and `th` are some of the most common pairs of letters in English). Except perhaps in proper names or words derived from other languages, some combinations do not occur at all, e.g. in English: `hj`, `bq`, `xz`. Similarly, some letters are often doubled (e.g. `ee`, `ff`, `ll`, `oo`, `ss`, `tt`) while others never are (e.g. `aa`, `ii`, `hh`, `jj`, `qq`, `vv`, `xx`). Furthermore, certain letters are more likely in certain positions, for example, in English, it is extremely rare to find more than three consonants in sequence; the letter `y` is more frequently found at the end of a word than at the beginning; and only a handful of words end in `i`[11].

The fact that words and letters are patterned in a structured, predictable manner was exploited by the well-known children's writer Charles Lutwidge Dodgson (a.k.a. Lewis Carrol), author of the "Alice in Wonderland" books. Not coincidentally, Dodgson was employed as a lecturer in mathematics at Christchurch Colleague, Oxford (specialising in the fields of geometry, mathematical logic, linear and matrix algebra, and recreational mathematics). One of Dodgson's / Car-

[11] The same is true of sounds: the distinctive `ng` sound / `ŋ` / as in thing, never occurs at the beginning of a word.

neare e delle matrici, e matematica ricreativa). Una delle creazioni più famose di Dodgson / Carroll è il poema "Jabberwocky" una parodia della letteratura in Inglese Antico, che appare in "Attraverso lo Specchio" (1871), e i cui primi versi sono riprodotti qui sotto[12]

> "Era brillosto, e gli alacridi tossi
> succhiellavano scabbi nel pantúle:
> Méstili eran tutti i paparossi,
> e strombavan musando i tartarocchi."

Questa è normalmente classificata come `letteratura nonsense`, ma si può vedere che fornisce qualche senso sebbene molte delle parole non sono affatto convenzionali in Italiano: ad esempio `brillosto`, `alacridi`, `tossi`, `succhiellavano`, `scabbi`, `pantúle`, `méstili`, `paparossi`, `strombavan`, `musando`, `tartarocchi`. Questo perché la grammatica e la sintassi sono ortodosse, in maniera tale che si può identificare senso (o funzione) grammaticale allo strano vocabolario: ad esempio `brillosto`, `alacridi`, `scabbi`, `méstili` sembra funzionino come aggetivi; `succhiellavano`, `strombavan`, `musando` sono verbi, e `tossi`, `pantúle`, `paparossi` e `tartarocchi` nomi. Il fenomeno psicologico detto `effort after meaning` [10] gioca un ruolo importante. Gli esseri umani hanno la tendenza nel ricercare senso in qualsiasi cosa porti un messaggio. Questa necessità fondamentale è visibile anche nel ben attestato fenomeno detto `pareidolia`, che si manifesta molto spesso in cose come vedere delle facce e oggetti in nuvole, o sentire delle parole in suoni casuali come nei fenomeni di voci elettroniche volgarizzato dagli `investigatori del paranormale`.

Un maggiore fattore di rassomiglianza del senso è che, sebbene parole come `brillosto` non esistano nel lessico italiano, sembrano parole italiane. Questo perché sono costruite con combinazioni di lettere come `br`, `ll`, `sl`, `osto` ecc. (si confronti ad esempio con parole come *enough* *which* o *thought* che non usano sequenze presenti in italiano). Infine, e molto ingegnoso da parte di Carrol, è la `fonoestesia` o il simbolismo fonetico che suggerisce quale possa essere il significato di qualche parola, e quindi implica la codifica ad

[12] Qui adotto la traduzione italiana intitolata "Il Ciarlestrone" di Adriana Crespi (1974): `http://www76.pair.com/keithlim/jabberwocky/translations/italian1.html` (N.d.T.)

roll's most famous creations was the poem "Jabberwocky", a parody of Old English literature, which appeared in "Through the Looking Glass" (1871), and whose first verse is reproduced below:

> "Twas brillig, and the slithy toves
> Did gyre and gimble in the wabe;
> All mimsy were the borogoves,
> And the mome raths outgrabe."

This is usually classed as `nonsense literature`, but it can be seen that it does convey some sense even though many of the words are not conventionally English at all: e.g. `brillig`, `slithy`, `toves`, `gyre`, `gimble`, `wabe`, `mimsy`, `borogoves`, `mone`, `raths`, `outgrabe`. This is because the grammar and syntax of the text is orthodox, so one can at least ascribe some grammatical sense (or function) to the strange vocabulary: for example `brillig`, `slithy`, `mimsy` and `mome` all appear to function as adjectives; `gyre`, `gimble` and `outgrabe` are verbs; and `toves`, `wabe`, `borogoves` and `raths` are nouns. The psychological phenomenon of what has been called the `effort after meaning` [10] also plays a part. Humans have a tendency to strive to see sense in anything that is presented as a message. This basic urge is also seen in well-attested phenomenon such as `pareidolia`, which often manifests itself in such things as seeing faces or objects in clouds, or hearing words in random sounds as in the Electronic Voice Phenomena popularised by "paranormal investigators".

A major factor in this semblance of sense is that, even though such words as `brillig` etc. do not exist in the lexicon of English, they do look like English words. This is because they are made up of possible letter combinations in English such as `br`, `ll`, `sl`, `ith`, `oves` etc. (compare for example, Italian words like *gnocchi*, *ahimé* or *sforza* which do not use English letter patterns). Finally, and most ingeniously on the part of Carroll perhaps, the `phonesthesia` or phonetic symbolism of some of the words themselves hints at what their meaning might be and thus entails encoding at a whole new level compared to what we discuss elsewhere in this article. For example, `slithy` is reminiscent of slither. `Mimsy`, with its

un completo nuovo livello confrontato con quello che abbiamo discusso in altre parti di questo articolo. Nel testo originario, `mimsy`, con il suono corto /ɪ/ (prodotto quando la lingua è alta e nella posizione avanzata della bocca, che crea una piccola cavità risonante che amplifica alcune alte frequenze) fa pensare, almeno in Inglese, a qualche cosa di piccolo o delicato (come *bit, little, twitter*). `Mome`, in contrasto con il suono /əʊ/ un dittongo, o combinazione del suono di due vocali, il secondo dei quali (/ʊ/) prodotto con la lingua abbassata nel retro della bocca (risultando in una grande cavità che amplifica le bassa frequenze) dà l'impressione di qualche cosa di più grande (ad esempio *row, tome, stone*) ([11], pag. 166-167).

L'analisi delle molte e varie sequenze che possono essere rilevate nel linguaggio naturale è diventata una parte importante della criptoanalisi. È anche un modo utile per distinguere testi realmente codificati da linguaggi di puro *nonsense*, pseudotesti, come quelli trasmessi da numerose stazioni utilizzate per confondere gli operatori e i criptonalisti del SIGINT (si veda la sez. 2.1). Forse il più noto esempio di questo esercizio nel distinguere senso e non-senso linguistico sono i tentativi di analizzare il cosidetto `Manoscritto di Voynich`. Battezzato così dal nome di un libraio polacco, è la pergamena di un codice illustrato risalente al XV secolo, come indica una datazione al radiocarbonio[13]. Comprende più di 170 mila caratteri scritti a mano in alfabeto e linguaggio sconosciuti. Questo fornisce ai criptoanalisti un enorme mole di lavoro. Nonostante questo, nessuno è ancora riuscito a decodificarne in maniera conclusiva, anche una minima parte. La teoria che sia una burla, contente un linguaggio privo di senso, è stata presentata da molti, ma il fatto che mostri caratteristiche simili a quelle di una lingua nella consistenza e frequenza nella sequenza di certe parole e lettere ha incoraggiato molti studiosi nel credere che si tratti di una genuina lingua. Negli anni '40 e '50 del secolo scorso, i criptoanalisti della `US National Security Agency` guidati dal leggendario William F. Friedman, ex capo del `US Signals Intelligence Services`

[13]Una riproduzione digitalizzata ad alta risoluzione dell'opera completa, conservata nella libreria Beinecke di libri rari della Università di Yale, può essere consultata al sito: `https://archive.org/details/voynich`

short /ɪ/ sound (produced when the tongue is high and at the front of the mouth, which creates a small resonant cavity that amplifies some higher frequencies) makes, in English at least, one think of something small or delicate (e.g. bit, little, twitter). `Mome`, by contrast, with its /əʊ/ sound, a diphthong (or combination of two vowels sounds), the second of which (/ʊ/) produced with the tongue low down to the back of the mouth (resulting in a large resonant cavity that amplifies low frequencies) gives the impression of something larger (e.g. crow, tome, stone) ([11], pag. 166-167).

Analysis of the many and diverse patterns that can be detected in natural language has become an important part of cryptanalysis. It is also a useful way of distinguishing between texts in real unknown codes or languages from real nonsense, pseudo-texts, such as those transmitted by some numbers stations used to confuse SIGINT operatives and cryptanalysts (see § 2.1). Perhaps the best-known example of such an exercise in telling the difference between linguistic sense and nonsense are the attempts at analysis of the `Voynich Manuscript`. Named after a Polish book dealer, this is an illustrated codex whose parchment originates, so radiocarbon dating shows, in the early 15^{th} century[12]. It comprises more than 170,000 handwritten characters in an unknown alphabet and language. It thus gives cryptanalysts a huge amount of material to work with. Despite this, no one has succeeded in decoding even a small portion of it conclusively. The theory that it is an elaborate hoax, constituting mere gibberish, has been put forward by many, but the fact that it does show language-like characteristics in its consistency and in the frequency of certain "words" and letter patterns has encouraged many scholars to believe that it is an example of a genuine language. In the 1940s and 50s, cryptanalysts working at the US National Security Agency led by the legendary William F. Friedman, former head of the US Signals Intelligence Services

[12]High resolution scans of the complete work, kept in the Yale University Beinecke Rare Book And Manuscript Library, can be viewed at: `https://archive.org/details/voynich`.

fallirono nel decifrarne anche una piccola parte, ma non riuscirono a scartarlo come una burla[14].

Frequenze e sequenze possono essere usate per distinguere tra ricorrenze casuali e significative in senso statistico[15], come nell'analisi dei segnali radio che arrivano da galassie lontane nella ricerca di vita extra-terrestre. Possono essere inerenti alla lingua e possono fornire importanti indizi ai criptoanalisti (almeno possono indicare in quale lingua sia scritto il testo originale cosa che sono stati incapaci di individuare con il codice `Voynich`). I criptografi si sono gradualmente resi conto di questo, e hanno concepito metodi in modo che queste sequenze, osservabili anche nel manoscritto di `Voynich`, possano essere mascherate. I princìpi soggiacenti nomenclatura e poligrafia possono essere rimodulati per questo scopo. Un notevole sviluppo fu un'invenzione, attribuita al Duca di Mantova, nel XV secolo, dell'omonimo codice dove le comuni lettere, o loro combinazioni, nel testo originale erano indicate da diversi caratteri o simboli nel testo cifrato secondo la loro frequenza, in questo modo, rendendoli più difficili da decifrare. Ad esempio, si può decidere di codificare la lettera `a` in due diverse maniere, sia come `d` o come £, e `t` come `w` e &. In questo modo, il nostro esempio "We attack at dawn", può diventare "zh dw&£fn dw gdzq" o "zh £&wdfn d& gdzq" e così via.

Nel XX secolo, i metodi di criptaggio implicano anche l'uso di diversi codici aleatori per ogni lettera o suono (o una loro combinazione) con cose

failed to decipher any of it, but could not dismiss it as a hoax[13].

Frequencies and patterns can be used to differentiate between random occurrences and significance (in a statistical sense)[14], as in the analysis of radio signals coming from distant galaxies in the search for extra-terrestrial life. They are also integral to language and can provide valuable clues to cryptanalysts (at least when they can guess which language the plaintext is in, which they have been unable to do with the `Voynich` manuscript). Cryptographers gradually realised this and devised ways so that such patterns, observable even in the `Voynich` manuscript, can be camouflaged. The principles underlying nomenclators and polygraphs could be repurposed for this purpose. A notable development was the invention, attributed to the Duca di Mantova, in the 15^{th} century of the homophone cipher whereby common letter or letter combinations in the plaintext were subtitled by different characters or symbols in the ciphertext according to their frequency, in this way, making them harder to identity. For example, one may decide to encode the letter `a` in two different ways, either as a `d` or a £, and `t` as `w` and &. In this way, our example "We attack at dawn", could become "zh dw&£fn dw gdzq" or "zh £&wdfn d& gdzq" and so on.

In the 20^{th} century, encryption methods also involved the use of different randomised ciphers for each letter or sound (or combinations thereof),

[14] Anche il famoso criptoanalista britannico John Tiltman (si veda la sez. 3.0) ha aiutato la squadra del NSA.

[15] Nella prima categoria possiamo classificare alcune discutibili applicazione della matematica nell'analisi di sequenze in testi come l'ampiamente pubblicizzato codice della Torah [12]. Si afferma di aver identificato un insieme di parole codificate nascoste nel libro ebraico della Torah secondo Drosnin [12] che messe insieme hanno predetto importanti eventi storici. Altri analisti, tuttavia, hanno messo in evidenza che scoperte simili a quelle di Drosnin possono essere ritrovate in qualunque testo molto lungo, e quindi i suoi risultati sono lontani dalla significatività statistica [13].

[13] The famous British cryptanalyst, John Tiltman (see § 3.0) also assisted the NSA team.

[14] And in the former category, we can class some rather dubious applications of mathematics analysis of patterns in texts, such as the much publicised Bible of Torah code [12]. This claims to have identified a set of encoded words hidden within the Hebrew text of the Torah according to Drosnin [12], put together, these words have predicted significant historical events. Other analysts however have pointed out that similar findings to those of Drosnin may be found in any long text, and thus his results are far from statistically significant [13].

come one-time pad[16], o macchine capaci di generare un numero quasi infinito di diverse combinazioni rendendo inutili analisi frequenziali fatte con carta e penna.

In queste sezioni (2.0-2.2), ci siamo limitati ad alcuni princìpi di base di codici relativamente semplici, quelli che esistevano prima del ventesimo secolo e l'avvento di congegni automatici di decifrazione ad alta velocità. Nella sezione 3.0, tratteremo uno dei codici più sofisticati ed importanti usati nella Seconda Guerra Mondiale allo scopo di illustrare come la matematica e la linguistica sono, insieme, state usate nella criptonanalisi.

3 Il cifrario di Lorentz (Tunny)

Al tempo della Seconda Guerra Mondiale, che fu la prima guerra in cui le forze armate si affidavano alle telecomunicazioni su grande distanza usando la radio, l'arte della cifratura si è evoluta in una scienza. Decifrare messaggi quindi diventò molto più difficile e al di là delle capacità di parolieri, poliglotti, entusiati enigmisti, giocatori di carte ed eccentrici assortiti che i militari avevano tradizionalmente trasformato in criptoanalisti[17].

In questa sezione, mi concentrerò su un esempio, quello che ha condotto alla costruzione di Colossus, menzionato nella sezione 2.0, e analizzerò il processo che ha condotto alle decifrazione del codice tedesco Lorentz (Tunny per gli alleati). Questo è meno noto di Enigma, lo spionaggio lo

[16] Il moderno sviluppo del cifrario di Vigenère (sez. 2.1), questo è un metodo ampiamente usato consistente una chiave fatta di stringhe di caratteri o simboli selezionati aleatoriamente che si combinano con le lettere nel messaggio originale con un'addizione modulata (come le ore del giorno). Se la chiave è: tenuta segreta; lunga almeno come il messaggio stesso (quindi evita le ripetizioni trovate nel cifrario di Vigenère); è usata solo una volta; ed è veramente aleatoria; si ritiene quindi indecifrabile.

[17] Tra questi c'era una grande proporzione di donne anche all'inizio del moderno spionaggio dei segnali. Nella Prima Guerra Mondiale, fino ad un terzo del personale impiegato nel britannico SIGNINT era femminile; questa percentuale è salita alla metà nella Seconda Guerra Mondiale. Le donne superarono in numero gli uomini in molti contesti lontani dalle linee del fronte [14]. A Bletchley Park il 75% dei criptoanalisti era composto da donne tra le quali spiccano figure come Mavis Batey, Joan Clarke, Jane Fawcett, e Jean Valentine [15, 16]

with things like one-time pads[15], or machines able to generate an almost infinite number of different combinations making frequency analysis of the pen and paper kind useless.

In these sections (§2.0-2.2), we have limited ourselves to some of the basic principles of relatively simple codes, those that existed before the twentieth century and the advent of high speed automated encryption devices. In Section 3.0, we will turn to one of most sophisticated and important codes used in WW2 as a way of illustrating how mathematics and linguistics were used in conjunction with each other in cryptanalysis.

3 The Lorenz (Tunny) cipher

By the time of WW2, which was the first war where armed forces were relying on telecommunications over vast distances using radio, the art of encryption had evolved into a science. Decrypting messages thus became much more difficult and beyond the ability of the sundry wordsmiths, polyglots, word puzzle enthusiasts, card players, and sundry eccentrics who the military had traditionally turned to for cryptanalysis.[16].

In this section, we will concentrate on one example, the same which led to the building of Colossus, mentioned in §2.0, and look at the processes leading to the cracking of one German cipher, Lorenz (Tunny to the Allies). This is less known than Enigma, and the intelligence labelled Ultra which it provided, but it was a far more complex code using quicker, more powerful en-

[15] A modern development of the Vigenère cipher (§ 2.1), this is a widely used method consisting of a key made up of a string of randomly selected characters or symbols which combine with letters in the original message by modulation addition (like hours in a day). If the key is: kept secret; is at least as long as the message itself (thus avoiding the repetition found in the Vigenère cipher); is only used once; and is truly random; then it is thought to be undecipherable.

[16] Among these were a high proportion of women even from the beginnings of modern signals intelligence. In WW1, up to a third of those engaged in British SIGNINT were female; by WW2, this had risen to a half. Women actually outnumbered men in many contexts away from the front line [14]. At Bletchley Park, 75% of the cryptanalysts were female among whom such key figures as Mavis Batey, Joan Clarke, Jane Fawcett, and Jean Valentine [15, 16]

ha etichettato Ultra ed era un codice molto più complesso che usava macchine di decifrazione più veloci e più potenti che richiedevano un minor numero di operatori (la macchina Enigma richiedeva, ad ogni terminale, tre operatori istruiti per inviare e ricevere.

Diversamente da Enigma (che usava un codice Morse) Tunny era basato su un sistema di telescriventi facile da usare che incorporava il codice di tipo binario delle telescriventi (anche noto come codice Baudot-Murray). Questo implicava la perforazione di un nastro di notevole spessore[18]. Il mittente digitava il testo non criptato e anche l'operatore dall'altra parte riceveva il messaggio non criptato. Dietro le quinte, c'erano due diversi strati di criptazione. La prima usava cinque ruote, e poi una seconda chiave era applicata ad altre cinque ruote. Due altre ruote generavano un "balbettio", introducevano irregolarmente una serie di caratteri che avrebbero introdotto un'aleatorietà apparente nella chiave[19]. Tunny era particolarmente difficile da gestire perché, a differenza di Enigma, nessuno, dalla parte degli alleati, aveva mai visto l'apparato che lo produceva, il Lorenz-Schlüßelzusatz SZ40/42, e questo avvenne solo al termine delle ostilità.

[18]Da questo il nome Tunny (pesce tonno): i britannici chiamarono con questo nome tutti i codici diversi dal Morse .

[19]Il numero totale di possibili configuazioni di criptazione che il Lorenz SZ42 avrebbe potuto generare era 1.6 x 10^{19} (23 x 26 x 29 x 31 x 37 x 41 x 43 x 47 x 51 x 53 x 59 x 61)

cryption machines that required fewer operators (the Enigma machine required three trained operators at each end to send and receive).

Unlike Engima (which used Morse Code), Tunny was based on a simple-to-use wireless teleprinter system that incorporated the binary-type code of the teleprinter (a.k.a. Baudot-Murray Code). This involved punching holes into ticker tape[17] the sender typed in plaintext and the operator at the other end received a plaintext message. Behind the scenes, it used two different layers of enciphering. The first cipher used five wheels and then a second key was applied using another five wheels. An additional two wheels would generate a "stutter", an irregularly inserted character that would introduce apparent randomness into the key [18]. Tunny was particularly difficult to deal with because, unlike the Enigma machine, no one on the Allied side had even seen the device that produced it, the Lorenz-Schlüßelzusatz SZ40/42, and were only to do so at the very end of hostilities.

[17]Hence, its codename Tunny (tuna fish): the British codenamed all non-Morse ciphers after types of fish.

[18]The total number of enciphering possibilities that the Lorenz SZ42 could generate was 1.6 x 10^{19} (23 x 26 x 29 x 31 x 37 x 41 x 43 x 47 x 51 x 53 x 59 x 61)

Testo originale		Carattere casuale		Testo cifrato		Carattere casuale		Testo originale
M		Q		J		Q		M
•	+	X	=	X	+	X	=	•
•	+	X	=	X	+	X	=	•
X	+	X	=	•	+	X	=	X
X	+	•	=	X	+	•	=	X
X	+	X	=	•	+	X	=	X

Tabella 1: *Processo di criptazione additivo per il cifrario Tunny.*
Basic Additive Encryption Process for Tunny cipher.

3.1 Progressi iniziali nel decifrare Tunny

Tunny usava un cifrario additivo, in cui una lettera casuale era aggiunta alla lettera del testo originale. Ad esempio, la lettera M nel testo originale (nel codice della telescrivente ••XXX) può essere combinata (per mezzo di una tavola della verità / una operazione XOR) con un carattere casuale Q (XXX•X), dando XX•X[20], che è il codice di telescrivente per J. La lettera M è quindi criptata come J nel testo cifrato. Per decifrare il testo, la macchina deve aggiungere lo stesso carattere casuale (Q) a J risultando ••XXX: il codice da telescrivente per M come nella tabella 1 [21].

Per la loro complessità le macchine Lorenz - Schlüßelzusatz SZ40/42, erano riservate ai più altri gradi del *Oberkommando der Wehrmacht* (OHW) Alto Comando Tedesco. Solo 30 erano in operazione (da confrontare con le circa mille macchine Enigma nelle varie versioni).

I messaggi intercettati venivano dati al leggendario decifratore e brillante linguista (poliglotta) brigadiere John Hessell Tiltman. Lui dedusse, correttamente, che Tunny era basato su una specie di codice Vennam (che usa un passaggio come operazione XOR come nella tabella 1). Inizialmente, affinché il ricevente possa sapere come regolare la sua macchina, il mittente deve trasmettere il settaggio della macchina con il messaggio (cosa che compromette la sicurezza). Questo fu fatto ponendo quello che i criptonalisti del Bletchley Park chiamarono un indicatore all'inizio della trasmissione. Dal 1942, quando Lorenz SZ40/42 divenne pienamente operativo, il sistema di indicatore consistette nella trasmissione in codice non criptato delle lettere "QEP" seguite da un numero di due cifre. Questo poteva essere identificato consultando un libro di codici noto ad entrambi gli operatori e mostrava il settaggio delle dodici ruote.

Infatti, quando due messaggi usano la stessa chiave, se sono addizionati o confrontati con un'operazione XOR (come nella tabella 1), allora gli effetti della chiave sono eliminati [17]. Que-

[20] Il carattere X significa che due simboli confrontati sono differenti; •significa che sono gli stessi: X + X o •+ •producono entrambi •e X + •o •+ X risulta in X.

[21] Adattata dal sito Bill Tutte Memorial Fund (https://billtuttememorial.org.uk/codebreaking/teleprinter-code)

3.1 Initial progress in cracking Tunny

Tunny used an additive cipher, whereby a random letter was added to the letter in the plaintext. For example, the letter M in the plaintext (in teleprinter code, ••XXX) can be combined (by means of a truth table / XOR operation) with the random character Q (XXX•X), giving XX•X [19], which is teleprinter code for J. M would thus be encrypted as J in the ciphertext. To decrypt the text, the machine would add the same random character (Q) to J resulting in ••XXX: the teleprinter code for M (as in table 1)[20].

Given their complexity, Lorenz - Schlüßelzusatz SZ40/42, machines were reserved for the upper echelons of the German *Oberkommando der Wehrmacht* (OHW) High Command. Only 30 were in operation (compared to thousands of Enigma machines of different versions).

Intercepted messages were given to the legendary codebreaker and brilliant linguist (polyglot) Brigadier John Hessell Tiltman. He correctly surmised that Tunny was based on a kind of Vennam cipher (that is, like a one-time pad employing a XOR operation, as in table 1). Initially, in order for the receiver to know how to set their machine, the sender had to transmit the machine settings within the message (which obviously compromised security). This was done by putting what Bletchley Park cryptanalysts called an indicator at the beginning of the transmission. From 1942, as Lorenz SZ40/42 came into full operation, the indicator system consisted of the transmission of the plaintext letters "QEP" followed by a two digit number. This could be looked up in a codebook issued to both operators and showed the settings of the twelve wheels.

Indeed, when two messages use the same key, if they are added together or compared in a XOR operation (as in table 1), then the effect of the key

[19] The character X means that the two symbols being compared are different; •means they are the same: X + X or •+ •would both result in •, and X + •or •+ X would result in X.

[20] Adapted from the Bill Tutte Memorial Fund website (https://billtuttememorial.org.uk/codebreaking/teleprinter-code)

sto significa che il risultato dell'addizione di due testi non cifrati dovrebbe essere lo stesso di due testi cifrati. In teoria, la chiave potrebbe essere recuperata dalla coppia dei messaggi cifrati e non. Questo è un processo lungo e laborioso se fatto a mano su ogni insieme di caratteri corrispondenti che sono stati generati dalla stessa chiave. Si può calcolare la somma dei caratteri cifrati (cioè il codice della telescrivente che corrisponderebbe all'operazione `XOR` dei due caratteri cifrati), ma non c'è un algoritmo matematico per decomporlo e trovare quale carattere del testo originale possa essere aggiunto per produrlo (allo stesso modo in cui non possiamo determinare in maniera univoca quali siano i due numeri da sommare per fare, diciamo 34: 14 + 20, 33 + 1, 9 + 25 etc.).

Tiltman, essendo un dotato paroliere di vecchia scuola, si affidava ad intuizioni linguistiche e perseveranza. Ad esempio, se la somma di due testi cifrati è `RSEZLS`, e si immagina che si riferisca ad una grande città della Gran Bretagna, allora, ricordando che `RSEZLS` è la somma di due testi non criptati si può fare una ragionevole ipotesi che una di loro sia, diciamo, London. Per verificarlo, si possono aggiungere i codici da telescriventi per `RSEZLS` e `LONDON` (come nella tabella 1) che produrrebbe `OXFORD`)[22]. In questo modo, si può essere sicuri che si sono identificati con successo i due testi non criptati dato che hanno senso compiuto nel contesto, ed è estremamente improbabile che sia il prodotto di un effetto casuale. Questi metodi si rivelarono efficaci, ma troppo lenti e avrebbero funzionato solo se si fosse in qualche modo dedotto il tipo di contenuto di ciò di cui si parlava.

Nell'agosto del 1941, i Britannici intercettarono due messaggi quasi identici inviati da un operatore negligente che fu forzato a rimandare un lungo messaggio di 4 mila caratteri. Nella sua fretta, o pigrizia, lui / lei non cambiò il settaggio della macchina `Lorenz SZ40/42`. Il fatto che ci fossero delle piccole differenze tra le due trasmissioni fu cruciale poiché due messaggi erano sicuramente meglio di uno. Due messaggi quasi identici fornivano `profondità informativa`:

[22] Questo esempio è dato da Frank Carter, a Bletchley Park Advisor, nel'eccellente documentario: Timewatch, Code-Breakers: Bletchley Park's Lost Heroes (BBC 2011).

is eliminated [17]. This means that the result of the addition of the two plaintexts should be the same as that of the two ciphertexts. In theory, the key can then be recovered from either ciphertext / plaintext pair. This is a long and laborious process if done by hand on every set of corresponding characters in two messages that have been generated by the same key. One is able to calculate the sum of the ciphertext characters (that is the teleprinter code that would correspond to the result of the `XOR` operation of the two ciphertext characters), but there is no mathematical way to decompose this and find out which plaintext characters could be added together to produce it (in the same way that what cannot determine mathematically which two numbers have been added together to make, say, 34: 14 + 20, 33 + 1, 9 + 25 etc.).

Tiltman, being an exceptionally gifted old - school wordsmith, relied on inspired guesswork, linguistic intuition, and perseverance. For example, if the sum of two ciphertexts is `RSEZLS`, and one suspects that each refers to a major British city, then, remembering that `RSEZLS` is the sum of the two plaintexts, one may make an educated guess that one of them is, say, London. To check this, one may add together the teleprinter codes for `RSEZLS` and `LONDON` (as in table 1), which would give `OXFORD`[21]. In this way, one could be fairly sure that one had successfully identified the two plaintexts, as they both made sense in the context, and were extremely unlikely to be a product of random chance. Such methods proved effective, but were slow and would only work if one could somehow deduce the kind of content one was dealing with.

In August 1941, the British intercepted two nearly identical messages sent by a sloppy operator who had been forced to resend a long 4,000-character message. In his / her haste, or out of laziness, he / she omitted to change the settings on the `Lorenz SZ40/42` machine. The fact that there were minor differences between the two transmissions was crucial as two identical texts would of course have been no better than one. It meant that the almost identical messages pro-

[21] This example is given by Frank Carter, a Bletchley Park Advisor, in the excellent documentary: Timewatch, Code-Breakers: Bletchley Park's Lost Heroes (BBC 2011).

costiuivano due diversi criptaggi generati dallo stesso settaggio (piuttosto che due possibilità di 1.6 miliardi di miliardi). In sintesi, fornivano una rara sequenza di 8 mila caratteri sulla quale Tiltman poteva lavorare.

Tilmann ipotizzò che il messaggio iniziasse con la frase SPRUCHNUMMER "numero del messaggio". Quando fu richiesto di ripetere, l'operatore impaziente lo accorciò in S P R U C H N R "messaggio no.". Questa piccola differenza significò che tutto ciò che seguiva la N nel secondo testo era diverso da quello della prima trasmissione. Tiltman continuò ad applicare la stessa tecnica, come illustrato nella tabella 1, usando la stessa chiave già adoperata per gli altri testi che fornivano la `profondità informativa`. Fu capace di verificare le sue intuizioni per ognuno dei testi dato che sapeva che criptavano, fondamentalmente, lo stesso messaggio. Questo gli dette la certezza che stava procedendo bene invece che trovare delle coincidenze accidentali. Tiltman non era un matematico e non aveva idea di come la macchina `Lorenz S40/42` potesse funzionare. Ci vollero 10 giorni per decifrare questi due messaggi (comunque una notevole impresa). Aveva perfezionato il metodo, ma non aveva la possibilità di ottenere una descrizione matematica della chiave che genera il processo. Se si fosse trovata, forse il processo di decifrazione si sarebbe velocizzato, o, addirittura, automatizzato.

3.2 Usare la matematica per ricostruire la macchina Lorenz S40/42

Tillman diede questo compito al nuovo arrivato, il ventiquattrenne laureato in chimica William "Bill" Tutte, che lavorò nella divisione di cui lui era il capo: la `Research Section`[23]. Come parte del suo addestramento come criptoanalista, a Tutte era stata insegnata la tecnica d'esame di Kasisnki, inventata da un ufficiale prussiano nel

[23] Umile di carattere, i talenti di Tutte non furono immediatamente evidenti. Fu esaminato da Alan Turing e bocciato come candidato per lavorare ad `Enigma`. Nella sua vita successiva, ottenne un PhD in Matematica da Cambridge e diventò professore all'Università di Waterloo nell'Ontario, dove contribuì a fondare il Dipartimento di Combinatoria e Ottimizzazione. Il suo ruolo nella decifrazione di `Tunny` non fu reso pubblico fino al 1990, molto dopo il suo pensionamento.

vided `depth`: they constituted two different encryptions generated by the same setting (rather than two randomly out of the 1.6 billion billion [sic] other possibilities). In short, they provided a rare 8,000 worth of characters with which Tiltman could work.

Tiltman guessed that the message began with the phrase SPRUCHNUMMER "message number". When asked to repeat, the impatient operator shortened this to S P R U C H N R "message no.". This tiny difference meant that everything form the first N in the second text was different from that in the first transmission. Tiltman was able to continue with the same additive technique, as illustrated in table 1, working through the ciphertext as he had with other `depths` that he had previously worked upon. This gave him the certainly that he was on the right track rather than just finding coincidental matches. Tiltman was no mathematician and had little idea of how the `Lorenz S40/42` machine might work. It took him 10 days to decrypt just these two messages (nonetheless a remarkable feat). He had however, perfected a method, but had no way to derive a mathematical description of the key generating process. If one could be found then perhaps the decryption process could be speeded up, perhaps even automated.

3.2 Using mathematics to reconstruct the Lorenz S40/42 machine

Tiltman gave this task to a newly-arrived 24 year-old chemistry graduate, William "Bill" Tutte, who worked in the unit he was head of: the `Research Section`[22]. As part of his training as a cryptanalyst, Tutte had been taught the Kasiski examination technique which had been invented by a nineteenth century Prussian officer to crack

[22] Humble in character, Tutte's talents were not immediately apparent. He had been interviewed by Alan Turing and rejected as a candidate to work on `Enigma`. In later life, he obtained a PhD in Mathematics from Cambridge and became a professor at Waterloo University, Ontario, where he helped found the Department of Combinatorics and Optimization. His role in cracking the `Tunny` was not public knowledge until the 1990s, long after his retirement.

XIX secolo per decifrare il codice di Vigenère (sez. 2.1). Questo metodo implica di cercare ripetizioni di stringhe di tre o più caratteri in un testo cifrato e misurare la distanza tra loro. È probabile che queste siano multipli della lunghezza della parola chiave, e, nel caso del cifrario Tunny permettono di invertire il processo che ha criptato il messaggio. Usando queso metodo laborioso, che era scomodo con lunghe sequenze che non potevano facilmente essere dispiegate su carta, Tutte, come Tiltman, ricorse a ragionevoli ipotesi. Dagli indicatori trasmessi con il messaggio (QEP più 2 cifre) fece la congettura che gli indicatori di Tunny fossero di 25 lettere (escludendo la J) per un settaggio e di 23 lettere per l'altro. Quindi, Tutte applicò la tecnica Kasiski sul primo impulso usando una ripetizione di 575 (cioè 25×23). Questo metodo non funzionò, eppure Tutte notò delle evidenti ripetizioni con un andamento non verticale, come previsto da questo metodo, ma leggermente diagonale. Quindi provò ancora con 574 con maggiore successo. I fattori primi di 574 sono 2, 7 e 41, quindi riprovò con una riga di 41 quadrati sul suo tavolo e riuscì ad osservare nelle colonne una moltitudine di sequenze ripetute indicando cha la sua intuizione era probabilmente corretta.

Questo fu solo il primo passo di un lungo percorso. Era chiaro che la macchina era formata da ruote separate, come Enigma, e Tutte dovette trovare quante fossero, e quale ruolo avessero nel processo di crittaggio. Fu capace di calcolare che c'erano due gruppi di cinque ruote ed altre due individuali.

Tutte indicò la prima componente della chiave come χ (chi) e immaginò che ci fosse un'altra componente che le era legata attraverso un'operazione XOR. Questo tuttavia non cambiava con ogni carattere (dato che le due ruote aggiuntive creavano un "balbettio"). Il risultato di questa ruota era detto ψ (psi). Trovò che questo assetto si applicava ad ognuno degli impulsi prodotti dai due gruppi di cinque ruote. Quindi, per un carattere individuale, la chiave era il risultato dell'operazione XOR di χ e ψ. La sequenza di caratteri aggiunti dalle ruote ψ era detto ψ esteso. Tutte riuscì a derivare le componenti di ψ grazie ad un difetto della macchina Lorenz S40/42 (corretto in seguito dai costruttori) che implicava che punti e croci fossero lievemente più proba-

Vigenère ciphers (§ 2.1). This method involves looking for repetition of strings of three or more characters in the ciphertext and measuring the distances between them. These are likely to be multiples of the length of the keyword, and, in the case of Tunny cipher, allow one to start to reverse engineer the way that it encrypted messages. Using this laborious method, which was cumbersome with longer sequences that could not be set out so easily on paper, Tutte, like Tiltman, resorted to educated guesses. He surmised from the indicators transmitted with the message (QEP plus 2 digits) that the Tunny indicators used 25 letters (excluding J) for one setting and 23 letters for the other. Tutte therefore tried a Kasiski examination on the first impulse of the key characters using a repetition of 575 (i.e. 25 $\times$ 23). This did not work but he did notice some obvious repetitions not on the vertical, as this method would have envisaged, but on a slight diagonal. He therefore tried again with 574, with more success. The prime factor of this number 574 are 2, 7 and 41, so he tried again with a row of 41 squares on his table and was able to observe in the columns a multitude of repeated patterns indicating that his intuition had probably been correct.

This was only the first step of many. It was clear that the machine consisted of separate wheels, like Engima, and Tutte had to set about finding out how many more there were, and what part they each played in the encryption process. He was able to calculate that there were two sets of five wheels and two more individual wheels.

Tutte called the first component of the key χ (chi) and speculated that there was another component which was linked to it via an XOR operation. This however did not always change with each character (because of the two additional wheels that created a "stutter"). He called the product of this wheel ψ (psi). This arrangement, he found, applied for each of the impulses produced by the sets of five wheels. So for a single character, the key was the product of the XOR operation of χ and ψ. The sequence of characters added by the ψ wheels was referred to as the extended ψ. Tutte was able to derive the ψ component because a flaw in the key setting of the Lorenz S40/42 machine (one later rectified by the manufacturers) meant that dots and

bilmente ripetuti (cioè seguiti da un altro punto o una croce). Avendo fatto questa scoperta, il resto della `Research Section` lo aiutò nello studio degli altri impulsi. Si trovò che le cinque ruote ψ si muovevano insieme sotto il controllo di due ruote - motore separate che chiamarono μ (mu).

In realtà, leggere un messaggio `Tunny` richiedeva non solo riuscire ad afferrare la struttura logica del sistema, come avevano fatto Tutte ed il resto della `Research Section`, ma implicava anche che si potesse derivare la sequenza cangiante delle camme (o dei *pin*) delle ruote[24]. Infine bisognava stabilire la sequenza iniziale (o la chiave del messaggio) del mescolatore delle ruote. La derivazione delle sequenze delle camme ("*wheel breaking*") era lo scopo della procedura chiamata `Turingery` dal suo inventore Alan Turing, che la propose nel 1942. Inutile dire che, insieme allo stabilire la sequenza iniziale, anche questo fu un altro compito molto difficile. Tutti questi passaggi furono inizialmente eseguiti laboriosamente a mano, ma avrebbero potuto eventualmente essere automatizzati, se gli alleati fossero stati capaci di leggere i messaggi `Tuny` abbastanza velocemente da capitalizzare il prezioso spionaggio che fornivano.

La storia del progetto e della costruzione di `Colosso`, il primo calcolatore elettronico programmabile, è un'altra storia (questa volta nel campo dell'ingegneria) del trionfo di conoscenza, intelletto, ingenuità e perseveranza contro le avversità. Max Newmann, il matematico, si rese conto che il il metodo di Tutte (una parte del processo di decriptazione di `Tunny`) poteva essere meccanizzato ma, avendo poca esperienza in ingegneria, la macchina che aveva costruito continuava a surriscaldarsi e a fermarsi. Tommy Flowers (un giovane ingegnere delle Poste che aveva una profonda conoscenza di valvole termoioniche e tubi a vuoto) fu coinvolto. Invece di aggiustare la macchina di Newton, progettò dall'inizio una macchina più efficiente e affidabile che funzionò

crosses were marginally more likely to be repeated (i.e. followed by another dot or cross). Having made this breakthrough, the rest of the `Research Section` could assist him in studying the other impulses. It was found that the five ψ wheels all moved together under the control of the two separate motor wheels, which they labelled μ (mu).

In fact, reading a `Tunny` message required not only that one had grasped the logical structure of the system, as Tutte and the rest of the `Research Section` had now done, it also demanded that one could derive the periodically changed pattern of active cams (or pins) on the wheels[23]. Finally, one had to establish the starting positions (or message key) of the scrambler wheels. The derivation of the cam patterns ("wheel breaking") was the target of a procedure called `Turingery` after its inventor Alan Turing, who came up with it in July 1942. Needless to say, together with the establishment of the starting settings, this was yet another highly complex process. All these different stages initially had to be done laboriously by hand, but could eventually be automated, indeed urgently needed to be done so if the Allies were to be able to read `Tunny` messages quickly enough to capitalise on the precious intelligence that they provided.

The story of the design and construction of the `Colossus`, the world's first programmable electronic computer, is another tale (this time in the field of engineering) of the triumph of insight, intellect, ingenuity, and sheer perseverance against incredible odds. Max Newman, the mathematician, had realised that Tutte's method (just one part of the process for cracking `Tunny`) could be mechanised but, with little experience of practical engineering, the machine that he built kept overheating and breaking down. Tommy Flowers (a young General Post Office telecoms engineer who had a thorough practical knowledge of things like valves and vacuum tubes) was brought in. Instead of fixing Newman's machine, he designed a more efficient and reliable one from

[24] Se incrementate producevano una `X` (1 in binario); decrementate generavano uno spazio (0 in binario) o • nella tabella 1.

[23] If raised, they generated an `X` (1 in binary); lowered, they generated a space (0 in binary) or •on table 1.

perfettamente la prima volta[25].

Senza dubbio lo sforzo valeva la pena. In un solo anno, gli Alleati conoscevano le intenzioni e le direttive dell'alto comando tedesco, a volte anche prima che fossero note ai comandanti di basso livello. Grazie a `Colosso`, i britannici riuscirono a passare ai Sovietici l'intero ordine di battaglia dei tedeschi per il loro attacco a Kursk nel Luglio 1943, senza rivelare i progressi che avevano fatto nella criptoanalisi[26]. Questa enorme battaglia, il più grande scontro di mezzi corazzati della storia, fu vinta dall'Unione Sovietica, e segnò l'inizio dell'inarrestabile avanzata dell'Armata Rossa su Berlino. La capacità di decifrare `Tunny` fu vitale anche nello sbarco del `D-Day`. Tra le altre cose, mostrò agli alleati che il massiccio piano per ingannare l'Asse sull'eventuale luogo dello sbarco ebbe successo. Lo spionaggio `Tunny` confermò che lo stesso Hitler era convinto, contro la più corretta opinione di alcuni suoi generali, che lo sbarco in Normandia del 6 Giugno fosse solo un diversivo.

4 Conclusione

La criptoanalisi è una delle aree dove la matematica è rilevante per la linguistica. È chiaramente un caso dove un approccio matematico può essere combinato con uno linguistico per produrre risultati che nessuno dei due campi produrrebbe da solo.

[25] L'esistenza di `Colossus`, ed il lavoro dei suoi creatori, Newman e Flowers fu mantenuto in stretto riserbo fino agli anni '70. Nel 1980, Flowers che, dopo la guerra, era ritornato a lavorare per le poste, ricevette un riconoscimento dovuto: fu il primo vincitore della medaglia Martlesham per i suoi risultati nell'informatica. Nel 1993 ricevette un semplice certificato in informatica come programmatore di *personal computer* dall'Hendon College (una scuola per adulti): questa è la sola qualifica in informatica che riuscì ad ottenere l'ingegnere responsabile del primo calcolatore elettronico programmabile.

[26] Il generale Zhukov, il comandante Sovietico, era convinto che i britannici avessero una talpa nel quartier generale dell'esercito tedesco (alimentando il sospetto che gli alleati occidentali potessero raggiungere un accordo di pace separato con i Nazisti), così impressionato con le informazioni che aveva ricevuto. Parte dei motivi per la segretezza nel lavoro di decodifica `Tunny` e dell'esistenza di `Colossus` poteva essere stata che i britannici sospettavano che nel mondo post-bellico i Sovietici (di cui Churchill diffidava sempre più) potessero cominciare ad usare `Lorenz`, o qualche cosa di analogo.

scratch. Amazingly, it worked perfectly first time[24].

Undoubtedly, all the effort was worth it. Within only a year, the Allies were learning about the intentions and directives of the German High Command, sometimes before Axis commanders lower down in the field were. Because of `Colossus`, the British were able to pass on the entire German order of battle for their attack on Kursk in July 1943 to the Soviets, without revealing the advances they had made in cryptanalysis[25]. This huge battle, the largest clash of armour in history, was won by the Soviet Union, and marked the beginning of the unstoppable Red Army advance on Berlin. The ability to crack `Tunny` ciphers was also vital in the `D-Day` landings. Among other things, it showed to the Allies that the massive plan to deceive the Axis about the location of the eventual landings, `Operation Fortitude`, had been successful. `Tunny` intelligence confirmed that Hitler himself was convinced, against some of his generals better judgement, that the landings in Normandy on 6th June were just a diversion.

4 Conclusion

Cryptanalysis is just one area where mathematics is relevant to linguistics. It clearly constitutes a case where a mathematical approach may be combined with a linguistic one to produce results which neither field by itself would produce.

[24] The existence of `Colossus`, and the work of its creators, Newman and Flowers was kept a tight secret until the 1970s. In 1980, Flowers, who had quietly returned to work for the GPO after the war, received some long overdue recognition: he was the first winner of the Martlesham Medal for his achievements in computing. In 1993, he received a simple certificate in basic information processing on a personal computer from Hendon College (a vocational school for adults): as it happened, the only formal qualification in computing that the engineer responsible for the first programmable electronic computer was ever to obtain.

[25] General Zhukov, the Russian commander, was convinced the British had a mole at the German army headquarters (further fuelling Soviet suspicions that the western Allies might reach a separate peace deal with the Nazis), so impressed was he with the information that he had received. Part of the reason for the secrecy over the work in deciphering `Tunny` and the existence of `Colossus` may have been that the British suspected that in a post war world, the Soviets (who Churchill increasingly distrusted) might start using `Lorenz`, or something based on it, themselves.

durre risultati che nessuno dei due campi isolati potrebbe produrre separatamente.

Oggi, con l'avvento della calcolo elettronico, annunciato da `Colossus`, e dell'Intelligenza Artificiale, è probabile che le applicazioni della matematica alla linguistica crescano esponenzialmente in molte direzioni differenti. La traduzione assistita da *Computer* è un'area dove sono avvenuti grandi progressi negli anni recenti; ci sono applicazioni gratuite *on-ine*, come `Google Translate`, capaci di fornire passabili traduzioni da e verso l'Inglese per molte lingue[27]. Questi programmi possono offrire la prospettiva di non dover imparare una lingua straniera, dato che sarà possibile usare i *computer* per parlare nella propria lingua ed essere istantaneamente capiti da chi parla qualsiasi altra lingua, analogamente a come `Lorenz SZ40/42` permetteva di trasformare un testo in un codice cifrato e quindi ancora nel testo originale. Ostler [18] sostiene che questa tecnologia eliminerà la necessità di una lingua franca globale, un ruolo attualmente goduto dall'Inglese:

> "È possibile guardare avanti nel mondo dei *media* elettronici interlinguistici, un mondo che si migliora e arricchisce dinamicamente. Come la rivoluzione della stampa, e altre varie rivoluzioni sociali associate all'urbanizzazione, hanno cambiato le regole di base della comunicazione tra gli europei del sedicesimo secolo, allo stesso modo, la tecnologia elettronica, se continua a seguire l'andamento attuale, è pronta per cambiare l'antica necessità di una lingua franca per tutti coloro che desiderano partecipare direttamente alle principali conversazioni internazionali. In breve, se l'elettronica può rimuovere la necessità di un intermediario umano per interpretare, o tradurre, le frustrazioni delle barriere linguistiche possono essere superate senza un mezzo universalmente condiviso al di là di un *software* compatibile. Discorsi registrati e

[27] Per l'accoppiamento di altre lingue, queste app potrebbero essere meno affidabili dato il minor numero di utenti e di dati di *input* con cui lavorare ed imparare.

Now with the advent of computing, as heralded by `Colossus`, and of Artificial Intelligence, the applications of mathematics to linguistics are likely to rise exponentially and in many different directions. Computer-Assisted Translation is an area where great progress has been made in recent years; even free online apps such as `Google Translate` are able to provide passable translations to and from many languages into English[26]. Such programs, may offer the prospect of no one having to learn foreign or second languages any more, as it will be possible to use computers to speak in one's own language and be instantly understood by speakers of any other language, analogously to the way that the `Lorenz SZ40/42` allowed plaintext to be transformed into ciphertext and then back again. Ostler [18] argues that such technology will do away with the need for a global lingua franca, a role currently enjoyed by English:

> "It is possible to look ahead into the dynamically improving, and enriching, world of interlingual electronic media. Just as the print revolution - and various other social revolutions associated with urbanization - changed the ground rules of communication among Europeans in the sixteenth century, so modern electronic technology, if it follows its current path, is set to change the ancient need for a single lingua-franca for all who wish to participate directly in the main international conversation. In brief, if electronics can remove the requirement for a human intermediary to interpret or translate, the frustrations of the language barrier may be overcome without any universal shared medium beyond compatible software. Recorded speeches and printed texts will become virtual media, accessible through whatever language the listener or speaker prefers.

[26] With pairings of other languages, they may be less reliable due to the smaller numbers of users and the less input they have to work with and learn from.

testi scritti diventeranno *media* virtuali, accessibili in ogni linguaggio che l'ascoltatore o il relatore preferisce.

Alla fine, e forse abbastanza presto, diciamo a metà del XXI secolo, ognuno potrà esprimere un'opinione nella propria lingua, sia oralmente o scrivendo, e il mondo comprenderà."

Ci si può chiedere se, dato che il linguaggio non è solo un mezzo per esprimere idee, ma anche una parte fondamentale della nostra identità e personalità, non sarà, nelle parole di Christiansen ([19], p.150),

> "dubbioso che gli esseri umani vorranno delegare la comunicazione con persone di altra lingua ad apparati digitali così come hanno entusiasticamente rinunciato all'aritmetica mentale per quella dei calcolatori elettronici. "

Tuttavia, il fatto che i traduttori automatici oggi siano così economici e facili da usare avrà certamente un effetto sul numero di persone che desiderano investire tempo, denaro e sforzo nell'imparare una seconda lingua straniera, che sarà una grande perdita per quanto riguarda i ben noti benefici psicologici e cognitivi di un individuo multilingua [20].

Un'altra area, con implicazioni indubitabilmente molto a più ampio raggio, è il programma di facilitazione linguistica tra umani e *computer* / *robot*, cioè i campi in rapida crescita del processamento del linguaggio naturale su cui lavorano molti informatici, ingegneri e matematici[28]. Il linguaggio parlato spontaneamente, con tutte le sue imperfezioni, false partenze e cambi di direzione, si è dimostrato difficile da processare per i *computer*, ma recentemente ci sono stati progressi enormi. Questi programmi funzionano perché le lingue possono essere analizzate e si possono scrivere algoritmi per permettere ai *computer* di

[28] Questa è un'area ovviamente popolata dai linguisti matematici, un ramo che applica specifici metodi e concetti matematici a sistemi linguistici, una tradizione, osserva qualcuno, che risale ad Euclide (circa 325 - 265 a.c.) e allo sudioso sanscrito Pāṇini, (circa 520-460 a.c.) [21].

> Ultimately, and perhaps before too long - say by the middle of the twenty-first century - everyone will be able to express an opinion in his or her own language, whether in speech or in writing, and the world will understand."

One may wonder whether, given that language is not only a means by which we express or ideas, but also a key part of our identity and our personality, it will not, in the words of Christiansen ([19], p.150), be

> "doubtful that human beings will want to delegate communication with people of other languages to digital devices as enthusiastically as they gave up mental arithmetic for electronic calculators."

However, the fact that useable translations are now so cheap and easy to come by will no doubt have an effect on the numbers of people willing to invest the time, money and effort into learning foreign or second languages, which would be a great loss given the well-attested cognitive and psychological benefits to an individual of being plurilingual [20].

Another area, with doubtless more far-reaching implications, is programs facilitating the linguistic or spoken interaction between humans and computers / robots, i.e. the rapidly expanding fields of natural language processing that many computer scientists, engineers, and mathematicians are working on[27]. Natural spontaneous spoken language, with all its imperfections and false starts and changes of direction, has proved difficult for computers to process, but recent years have seen enormous advances. These programs function because language can be analysed and algorithms written to allow computers to use, or

[27] This is an area covered most obviously by mathematical linguistics, a branch which specifically applies mathematical methods and concepts to linguistic systems, a tradition, some observe, going back to Euclid (c. 325-265 BCE) and Pāṇini, the Sanskrit scholar, (c. 520-460 BCE) [21].

usare il linguaggio, o di simularne l'uso[29]. La prospettiva che gli essere umani possano usare il loro linguaggio naturale come interfaccia con *computer* ed intelligenze artificiali (o tra *computer* e intelligenze artificiali senza alcuna intermediazione umana) è qualcosa che la fantascienza ha ipotizzato già da molto tempo. È un'ipotesi stuzzicante con cose come `Siri` di Apple o `Alexa` di Amazon, sebbene sia difficile prevedere quando diventerà realtà, ed esseri umani e *computer* potranno avere lunghe, e significative, conversazioni come avviene, ad esempio, nei film come `2001: Odissea nello spazio` di Kubrick, con HAL 9000 l'inizialmente geniale *computer* di controllo della nave spaziale `Discovery One`. Per ottenere questo livello, come nel caso della decifrazione del cifrario `Lorentz` (`Tunny`), linguisti, matematici, ingegneri, ed informatici dovranno lavorare insieme, poiché, separatamente nessuno di loro sarebbe capace di superare la miriade di sfide da affrontare.

Il fatto che matematici, linguisti, e studiosi di altri campi possano collaborare in maniera proficua non è una sorpresa. È semplicemente il risultato di una profonda connessione tra linguaggio, matematica e altre scienze; ognuna è una manifestazione di diversi ambiti cognitivi la cui funzione è quella di aiutare a comprendere aspetti differenti del mondo, e che forniscono diversi strumenti concettuali e affrontano lo stesso problema da prospettive diverse.

[29] Così è come Alan Turing presentò il *test* chiamato con il suo nome: non come se una macchina potesse usare il linguaggio come un essere umano (forse impossibile da verificare visto che il lavoro della mente e la locazione del linguaggio sono ancora coperti dal mistero) ma piuttosto se un essere umano sarebbe capace di distinguere se sta parlando con un *computer* o con un altro essere umano.

at least appear to use language[28]. The prospect of humans being able to use their natural language as an interface with computers and artificially intelligent devices (or indeed between computers and AI devices without any human agency) is something that science fiction has long assumed would one day be possible. Such a thing seems tantalisingly close with things like Apple's `Siri` or Amazon's `Alexa`, although it is difficult to predict when it will become reality, and humans and computers will be able to have long, meaningful conversations of the kind seen in films such as Kubrick's 2001: A Space Odyssey (1968) with HAL 9000 the initially genial control computer on the spacecraft Discovery One. To achieve this, as was the case with the cracking of the `Lorenz` cipher (`Tunny`), linguists, mathematicians, engineers, and computer scientists will have to work together, each looking at different aspects of the problem, because, separately, none of them will be able to overcome the myriad challenges.

The fact that mathematicians, linguists, and scholars from other fields can profitably work together should come as no surprise. It is merely a result of the underlying connection that exists between language, mathematics and the other sciences; they are each manifestations of different cognitive frameworks whose function is to help us understand different aspects of the world, and which give one different conceptual tools and thus approach the same problem from different perspectives.

[28] This, in essence, is how Alan Turing presented the problem when he thought up the test named after him: not as such whether a machine could actually use language in a humanlike way (something perhaps impossible to test seeing that the inner workings of the mind and language's place within it are still shrouded in mystery) but rather whether a human would be able to tell whether they were conversing with another human or not.

[1] N. Chomsky: *Cartesian linguistics: a chapter in the history of rationalist thought*, Harper & Row, New York (NY) (1966).

[2] C. R. Darwin: *The descent of man, and selection in relation to sex*, John Murray, London (1871).

[3] S. Pinker: *The Blank Slate: The Modern Denial of Human Nature*, Viking, New York (NY) (2002/2016).

[4] N. Chomsky: *Lectures on Government and Binding*, Foris, Dordrecht (1981).

[5] F. De Saussure: *Cours de Linguistique Générale*, Payot, Paris (1916).

[6] N. Chomsky: *Aspects of the Theory of Syntax*, Massachusetts Institute of Technology Press, Cambridge (MA) (1965).

[7] N. Chomsky: *Language and Mind*, Cambridge University Press, Cambridge (2006).

[8] G. H. Hardy: *A Mathematician's Apology,* Cambridge University Press, Cambridge (1940).

[9] I. A. Al-Kadi: *The origins of cryptology: The Arab contributions,* Cryptologia, 16 (1992) 97.

[10] F. C. Barlett: *Remembering,* Cambridge University Press, Cambridge (1932).

[11] S. Pinker: *The Language Instinct,* Penguin, Harmondsworth (1994).

[12] M. Drosnin: *The Bible Code,* Simon & Schuster, New York (NY) (1997).

[13] B. McKay, D. Bar-Natan, M. Bar-Hillel, G. Kalai: *Solving the Bible Code Puzzle,* Statistical Science, 14 (1999) 150.

[14] J. Ferris: *Behind the Enigma: The Authorized History of GCHQ, Britain's Secret Cyber-Intelligence Agency,* Bloomsbury Publishing, London (2020).

[15] T. Dunlop: *The Bletchley Girls,* Hodder & Stoughton, Lomdon (2015).

[16] M. Smith: *The Debs of Bletchley Park,* Aurum, London (2015).

[17] W. T. Tutte: *Appendix 4: My Work at Bletchley Park*, in B. J. Copeland (ed.) *Colossus: The Secrets of Bletchley Park's Codebreaking Computers* Oxford University Press, Oxford, (2006), pp. 352-369.

[18] N. Ostler: *The Last Lingua Franca: English until the Return of Babel,* Penguin, London (2010).

[19] T. Christiansen: *The rise of English as the global lingua franca. Is the world heading towards greater monolingualism or new forms of plurilingualism?,* Lingue e Linguaggi, 15 (2015) 129.

[20] D. Mehmedbegovic, T. H. Bak: *Towards an interdisciplinary lifetime approach to multilingualism,* European Journal of Language Policy, 9 (2017) 150.

[21] A. Kornai: *Mathematical Linguistics.,* Springer, Berlin (2008).

Thomas Christiansen: Thomas Christiansen è professore associato di Lingua e Traduzione, per la Lingua Inglese presso l'Università del Salento e, dal 2016, Direttore del Centro Linguistico di Ateneo. Ha insegnato in varie università della Puglia, del Regno Unito e della Polonia. Ha completato il dottorato di ricerca in linguistica testuale a Salford (UK). Ha svolto ricerche in varie aree della linguistica e ha pubblicato monografie e articoli su molti campi, tra cui linguistica sistemica e grammatica funzionale, varietà di inglese, inglese come lingua franca, didattica della lingua inglese, valutazione delle competenze linguistiche, e analisi di diversi corpora, incluso il discorso parlato. Ha anche lavorato per molti anni come consulente esperto per Cambridge Assessment English.

Thomas Christiansen: Thomas Christiansen is associate professor in English Language and Translation at the Università del Salento (Lecce, Italy) and since 2016, Director of the University Language Centre. He has taught in various positions at various universities in Apulia (Italy), the UK, and Poland. He completed his PhD in textual linguistics at Salford (UK). He has researched into various areas of linguistics and published books and articles on many fields including systemic linguistics and functional grammar, varieties of English, ELF, teaching English, language testing, and analysis of different corpora, including spoken discourse. He has also worked as an expert consultant for Cambridge Assessment English for many years.