Languages and Knowledge

Artificial Intelligence and Machine Translation in Scientific Communication
1- *SCIENCE AND THE LANGUAGE BARRIER*

**LINGUA FRANCAS AS A REMEDY**
Science is a sophisticated human creation: The language needed to express it does not arise spontaneously, it is the result of a great effort.

• In order to meet the concepts of science, languages have to develop a specific vocabulary by adapting words and sentences to its specific epistemic needs.
WORDS AND TERMS

To study science is to study a new language where words as new as *quark* or as everyday as *cell, nucleus, energy and work* take on a precise semantic value that is different from the usual one.

*In science, words become terms with an unambiguously defined and universally shared meaning.*
Each language has its own structure and history. This makes the transmission of ideas between speakers of different origins problematic.

Historically this barrier has been mitigated by giving a language, known as lingua franca, a supremacy as a means of international communication between people with different mother tongues, to whom this language is foreign.

A lingua franca is a language deliberately acquired for reasons of utility to facilitate communication between speakers of different mother tongues.
THE RISE AND Transience Of LINGUA FRANCAS

There are three main factors that favour the transformation of a language into a lingua franca: the creation of an empire, trade and missionary religions.

Lingua francas do not last indefinitely: the loss of the historical, military, commercial or religious causes which decreed their predominance finally decree their decline too.

A lingua franca is a language of convenience. When it ceases to be convenient - however widespread it has been - it will be dropped, without ceremony, and with little emotion.
Latin was the language of science for a few centuries.

It was the lingua franca in which the scientific revolution of the Baroque age was expressed.

Galileo and Kepler wrote in Latin, and Newton entitled *Philosophiae Naturalis Principia Mathematica* the work in which he described the law of universal gravitation (1687).
THE (COUNTER)REVOLUTION OF NATIONAL LANGUAGES

The 18th century, which was the century of the rationalisation of scientific language, was also the time when Latin was abandoned.

Latin survived as the matrix of specialist nomenclature, but was used increasingly rarely as a language of communication: scholars turned to national languages, and science spoke French, English, German, Italian and Swedish.
Europe's academic elites nevertheless retained a fairly solid knowledge of Latin for at least two centuries.

*Scientific language, developed on this shared core of terminology, seemed to be able to be expressed in all the scientific languages of the time with the same adequacy and precision.*
INDUSTRIAL CIVILISATION AND THE RETURN TO BABEL

The illusion of universality dissolved in the course of the 20th century, when the rise of industrial civilisation raised the number of scholars, expanded the tasks of science and moved its boundaries far beyond Europe.

**Great science** was born with its large-scale projects, massive funding, complex equipment and extensive laboratories.

The weight of technological *spin-offs* transformed the rules governing the dissemination and application of scientific knowledge:

**By changing the world, science also changed the conditions of its own development.**
THE MULTIPLICITY OF LANGUAGES: AN EPISTEMIC RISK?

Faced with the enormous wealth of knowledge generated by a firmly international collaboration, scholars began to perceive the multiplicity of languages as dangerous for the epistemic norms of 'open science'.

Those who publish in a minor language, which the majority of people does not master, risk shattering the unity of the global scientific system because they remove their claims from peer review.
HISTORY DECIDED ON ENGLISH

The political and military history of the West decided for all, and English became the vehicular language of science.

After a few decades, the *Economist* (December 1996) stated: "*English has gained an unassailable position as the world's standard language: it has become an intrinsic part of the global communication revolution*."

*In the age of the knowledge economy, English is not only the technical core of scientific language, but also the language used to talk and argue about science.*
The wealth of knowledge accumulated over decades of use has made English a valuable tool for the international communication of science.

*Recent profound changes in political and economic balance may favour the competitive spread of other languages;* yet so far none seems to have had the strength to replace English as the main lingua franca of science.

*Risks to the English language do not come from history but from technological progress.*
In 2010 a very well-documented and unexpected book proposed that lingua francas would be supplanted *not by military, commercial or religious events but by technological innovations. Rapidly advanced machine translation would make any lingua franca useless.*

*English would become the last lingua franca of human history.*
THE HYPOTHESIS OF A MULTILINGUAL WORLD WITH MONOLINGUAL INDIVIDUALS

In Osler's model, the future will be multilingual to a much more radical extent than in the past, but only because no one will have a practical need to learn foreign languages.

As universal English becomes a mere memory, everyone will use words from the easiest language for them, without having to worry about the language of their listeners.

The world will be comprehensible in all its diversity.

(N.Osler)
2- THE TECHNOLOGICAL REVOLUTION AND ARTIFICIAL INTELLIGENCE

THE NEW PROTAGONISTS OF LINGUISTIC INTERCOMPREHENSION
Automatic language processing, usually called NLP (Natural Language Processing), is one of the most complicated branches in the domain of Artificial Intelligence.

For a number of years computer scientists have attempted a rationalistic approach to language under the illusion of instructing machines to understand words semantically.

The project proved to be unproductive, and was replaced by an empirical approach that prescinds from machine understanding and is known as distributional semantics.

This is the system that seems to work well in many applications.
AUTOMATIC LEARNING (Machine learning)

Recent advances in Machine Learning using Deep Neural Networks have suggested that the problem could be solved by providing machines with sufficient volumes of data (terabytes).

The NPL method has been the key to the success of AI projects dedicated to language.

It is thanks to these systems that considerable progress has been made in the field of machine translation in recent years.
THE DISTRIBUTIONAL HYPOTHESIS USES TERABYTES OF TEXTS

The 'neuronal' NPL method is based on a strategy centered on avoiding the problem of word meaning. It adopts the distributional hypothesis, and states that there is a relationship between the meaning of words and the way they are distributed (i.e. occur and combine) in texts.

Since "words similar in meaning appear in the same contexts", it can be assumed that distributional semantics can serve as a basis for applying Machine Learning techniques.

The network provides us with terabytes of text, and it is therefore relatively easy to compute the distribution and combinations of words.

Approximating the meaning of a word to its distribution is mathematically convenient.
SOME REFLECTIONS ON THE DISTRIBUTIONAL HYPOTHESIS

Distributional semantics has an empirical basis, but works well in many applications.

It allows the assessment of similarity of meaning between words and sentences with greater accuracy, robustness and flexibility than could be achieved using synonym vocabularies or other lexicographic resources.

This capability is crucial for many Machine Learning applications, such as text classifiers or automatic translators.

For this type of applications it is not necessary to know what words mean; it is sufficient to have a measure of their similarity due to the fact that they occur in the same sentences (e.g. stroking the dog/cat).

With a lot of good data and with the necessary adjustments distribution matrices today work better than any dictionary in many practical cases.
Advances in machine translation are now in everyone's experience. Many mobile phones enable the reading of a scientific text or a newspaper in many languages other than the original. Unfortunately, as the example shows, Italian is still missing.
An important innovation is the creation of a free online supplement to NATURE magazine. Opposite is the text of one of the first issues, dedicated to the fight against the Coronavirus pandemic in Italian laboratories. As shown in the red box, you can just click on 'read in Italian', and the text will immediately appear in Italian language.
I laboratori italiani si reinventano contro la pandemia

Il desiderio di contribuire e nuovi finanziamenti spingono scienziati dai campi più diversi a riorientare competenze e attrezzature per studiare il COVID-19

Giorgia Guglielmi
Thank-you