Same Same but Different: Paradigms in Syntax

Kim Gerdes

Université Paris Nanterre

Sciences du langage

Traitement automatique des langues

Mémoire de synthèse en vue de l'obtention de l'Habilitation à diriger des recherches

Soutenance le 29/11/2018

Jury :

Prof. Ryan Abbott, University of Surrey and UCLA

Prof. Eva Hajičová, Univerzita Karlova, Prague

Prof. Sylvain Kahane (référent), Université Paris Nanterre

Thierry Poibeau, CNRS and ENS, Paris

Benoît Sagot, Inria, Paris

Prof. Irène Tamba, EHESS, Paris

Prof. Leo Wanner, Universtitat Pompeu Fabra, Barcelona

Table of Contents

Same Same but Different[1]: Paradigms in Syntax

1 Introduction

La mort [...] n'est pas drôle quand-même […]
parce qu'elle ne supporte pas la répétition[2]

Boris Vian

Repetition has a bad reputation. And yet, it is omnipresent in language production, not only in spoken language, where the speaker often approximates the desired description by juxtaposing almost synonymous phrases with few additions, thus leaving to the speaker the task of combining the reformulations into a coherent description (Blanche-Benveniste 1990). I will argue that language itself is mainly about repeating – repeating the right thing at the right moment. Language does not give you freedom, it obliges you to say things you do not actually want to say (Boas 1938, Jakobson 1959). Mastering a language is using the right collocations, it is using the combinations that many others have used before. It is about repeating, about not surprising anyone, about reducing entropy.

By placing the questions that I have asked myself during my 20 years of linguistic research into a timeline of linguistics as a whole, I hope to show that there actually is a unifying thread, the *fil conducteur*, the *leitmotiv* of my work. That is why this text is also about paradigms and paradigm changes. Two completely different types of paradigms: Syntactic paradigms and scientific paradigms. Right from the start of my PhD thesis, I have worked on loads of similar sentences, paradigms of sentences. And the questions that I have asked myself are not independent of scientific paradigms in the sense of Kuhn (1962): The context of what can and must be asked in Science.

This is a cumulative habilitation. So, instead of saying the things that I have written before a little differently, I try here to fill in the blanks, the missing links between the texts, while referring the reader to the original papers in the annex of this habilitation. All references to my own work are in bold face.

Another reason for this choice is that I like the word "cumulative". I believe in cumulative science with my positivist belief in the unique Truth that can only be found cooperatively by sharing access to knowledge, i.e. I believe in open access and free resources. I find it an amazing fact that the basic resource of today's society, knowledge, is no longer scarce but can be reproduced without significant cost. Without falling into the false hope that this will overcome the digital or other

1 The title refers to the Thai-English expression used to describe subtle nuances. Its usage is exemplified by this dialogue:
Q: *Is this a real Rolex?*
A: *Yes Sir, same same but different.*
From https://www.urbandictionary.com/define.php?term=same%20same%20but%20different

2 'Death just isn't funny because it can't stand repetition.'

divides, I do believe that the possibility to share knowledge freely is the determining feature of our present-day society.

However, I will try to convince the reader that knowledge is scarce in a different and unexpected way: There simply are not indefinitely many different ideas. For basic cognitive, combinatorial, and linguistic reasons, there cannot be an infinity of ideas, at least not in the human-brain centered sense in which we use the word today.

First, allow me to take you on a quick and biased ride of (my whig of) the *History of Science*, where I try to argue that all fields of science are conditioned by human and societal needs, desires, and limitations. I am particularly interested in the question of whether good answers to scientific questions have to be simple. This is followed by a succinct *History of linguistics* in Chapter 3, where I give an overview of different historic and present-day approaches to linguistics. I attempt to show how wrong scientific expectations lead to wrong linguistic questions, and then how sociological questions brought us to important branches of linguistics losing all contact to empirical research. With this background, I can finally turn to my own scientific pathway: Why questions that seemed relevant to me do not seem central anymore – partly because I have understood them all too well, and partly because I know now that the questions asked were based on wrong, simplistic preconceptions. Chapter 4 has essentially two segments: on Tree Adjoining Grammars and on topological descriptions of word order. Then follows the description of my ongoing work on corpora constitutes the central chapter of this text: After some brief remarks on corpus linguistics, I narrate my path to empirical data via my understanding of statistical analysis and via the tools and annotation schemes I have developed for spoken and written texts. I describe some ways I see how to gain knowledge from the large corpora that are now at our disposal, via detection of correlation and by exploring Machine Learning as a linguistic tool. Chapter 6 is devoted to Machine Creativity and how semantic coverage can lead to an *exhaustion of language*. I will conclude with some bold predictions on the linguistics of the future.

2 On the history of Science

2.1 Science's place in society

A science such as linguistics does not operate in empty space. Science answers questions asked by the scientist's society; it satisfies needs of the scientist's society, answering questions sometimes stemming from engineering. Furthermore, the raison d'être of science is often to give reason and rationale, even some kind of repentance, to the society's bad habits and unjustified preconceptions. Science is prestigious because it is perceived as impartial and, more importantly, essential for controlling the world, politically and intellectually. In my eyes, this advantage is not necessary to justify the existence of science itself – science is a much deeper human necessity as I will argue below – but Science's political benefits are quintessential in the public discourse on spending money on science and education.

General education, education beyond the lucky few, was not born out of philanthropy. A society with compulsory education had a competitive industrial and military advantage. Some illustrious 19th century contemporaries of the introduction of compulsory education fought it vigorously on grounds of the dangers that it brings to the ruling class: "If you want slaves, you are a fool to educate them to be masters."[3] The idea was that the Übermensch can only thrive if he has slaves. And making the plebs autonomous by leading them out of their "self-incurred immaturity" will deprive the best people of their slaves, forging a mediocre society. In Section 3.4 , I will return to the effects of massification of the humanities on linguistics.

The rise of smart machines, however, terminates the contract between science and the rest of society: We do not need educated masses anymore. A few can build and control smart machines to run every necessary task of the whole society. This schism goes much deeper than questions of financing science. It may simply no longer be true that understanding a phenomenon by modeling it is a precondition for building a good machine. Machine learning divides engineering and science. A good data scientist can build predictive models of societal phenomena such as petty crime or religious changes (Gore et al. 2018) just as well as building predictive models of physical phenomena such as a nuclear decomposition processes – all with the same technology, without the need for much specific domain knowledge. And of course the same technique can build syntactic parsers, machine translation systems, or even complete natural language understanding systems that do away with the need of any intermediate representation between meaning and text, leaving linguistics as a big black box filled with a neural network that humans find hard to take as an answer.

Then, and I shall return to that point in the concluding Chapter 7, all science may become human science in the sense that it will no longer serve engineering but rather human satisfaction. Good scientific explanations can then only be distinguished from bad scientific explanations by the degree of pleasure they trigger in the human that comprehends the explanation. Just as the best way to measure the quality of two ice-cream flavors can only be measured by the pleasure it causes, not by

3 "Will man Sklaven, so ist man ein Narr, wenn man sie zu Herren erzieht" Nietzsche, Werke, Band VIII, p. 153.

a greater benefit of one flavor over the other in any applied or independently gaugeable sense.

The problem with measuring pleasure is that human preferences are not necessarily reflections of what is actually good for the human (Ehrenfeld 1981). This describes the crisis faced by humanistic values as the world's dominating ideology, as Yuval Noah Harari argues convincingly in his *Homo Deus* (2016). Put simply, it is clearly hard to affirm in times of the Trump regime that humans know what is good for themselves. Maybe Big Data does know better after all? Maybe a statistical machine translation system does understand and describe language better than a linguist. It measurably works better than the translation rules that humans have come up with. Maybe it also constitutes a better explanation?

2.2 Darwin and the limits of cognition

In this section, I want to argue that our limited cognitive capacities made us look for low-level answers to many questions – to scientific as well as to moral questions. Our mind has evolved for down-to-earth tasks: for running, foraging, socializing – for surviving in the steppes. Not necessarily for looking for the right answer.

Darwin's idea, as any scientific idea, did not appear out of nothing. Most ships of the European colonizers brought researchers such as Darwin in order to understand what was there to see – and to steal – in the new territories, how to make use of the new continents with their plants, animals, and humans. Yet, when Darwin's ideas seeped in the discussions of the cultural salons of the 19th century, it was upsetting to accept many of the implications of the discovery of evolution.

On the one hand the theory of evolution was a logical continuation of humanistic ideas, originally from Stoicism[4], taken up by Christianity and the Renaissance, by including more and more beings inside worthiness, in the realm of rights, in the égalité: thy neighbor, the residents of the next valley, inhabitants of the next continent, slaves, women, and, finally and ongoing: other non-human species. Yet, the submission of the "New" World needed moral justification – and the idea of the survival of the fittest went down well with slavery and colonialism.

On the other hand it kicked humans off the high horse of their exceptionalism. Not only are apes our close relatives – we are a younger sub-species, which is why we should be called *Pan*[5] *sapiens* rather than *Homo sapiens*. It has started a still ongoing process of discovering the inadequacy of our reflexes to the modern times and more generally the limitations of our brain.

2.3 Understanding the human mind's imperfection

If humans are adapted to the steppes, even agriculture is a modernism that humans are not set up to deal with successfully. The innumerable victims of diseases, famines, and wars that have brought

4 *"For example, the second century Stoic Hierocles posited an early form of "cosmopolitanism," whereby the ego, the 'I' at the centre of our ethical life, was enfolded by concentric circles of moral concern. The closer the circle to the centre, to the I, the more demanding on our affections its subjects tend to be. The family was closest, and eventually one would reach all humankind in the outermost circle. For Hierocles, the goal of ethics encompassed the gradual extension of our sympathies to ever more distant circles of concern."* (Bassett 2018)

5 *Pan* is the taxonomical genus for the common chimpanzee and the bonobo, given in reference to the Greek god of nature and wilderness, who is depicted as a hybrid of human and animal.

mankind more than once to the brink of extinction are witness of the dangers of this “new” agricultural lifestyle. Even much more recent is the industrialization of agriculture and the over-availability of calories to most human beings on the planet. Today’s obesity epidemic shows mankind’s non-adaptation to the new situation. There is a scientific explanation for this phenomenon, scientific in the sense that it fits the data in a simple and greater picture: Evolution programmed us to fancy and overeat on fat and calories if they are available because this allowed us to survive long periods of starving. Those who did not overeat did not survive and their genes did not make it into the next generation. So any medical, psychological or educative solution to obesity would do well to take into account the evolutionary context.

Yet another example of evolutionary explanations to social phenomena concerns superstition and religious faith. At first view, it seems astonishing that counterfactual beliefs have not been weeded out by evolution. Imposed limitations in food and partners, which are common restrictions inside religious communities, as well as fear of non-existent monsters should be an evolutionary disadvantage. Again, to really understand this human behavior, we have to look beyond cultural traditions. We have to turn to evolution. One proposed solution is that it has been an enormous advantage to blindly trust the beliefs of one’s parents (Dawkins 2006). Children who did not believe their parents and just gulped down the obscure mushroom, very often did not make it long enough to produce offspring themselves.

This explains how wrong memes[6] can survive, but it does not explain how the idea evolved in the first place. Believing in omnipotent gods may be reassuring or scary as a thought, but the main reason may lie in our inborn logic: One key success of our species is the capacity to discover cause and effect relations. We understand how fire can be made by flint stones, how gravity makes water go downhill and not up, and we understand other humans’ motivation of their acts.

Humans are particularly good at recognizing faces. Only very recently machines have been able to catch up (although, and we will come back to that in Section 5.8 , that does not mean that we have genuinely understood how the face recognition system of the human brain actually works). This constantly running system is well-performing but regularly overheating. This is called *pareidolia*: seeing faces in clouds or maps – seeing faces where clearly there are none. More generally, we speak about apophenia (from ἀπό, ‘coming from’ and φαίνω, ‘to shine’, meaning ‘make appear’ or ‘discover’) an over-interpretation of actual sensory perceptions. The phenomenon, also known as overfitting in the world of machine learning, is more present among artists – and among schizophrenic patients. People and machines that see connections, false or real, that others do not

6 The term *meme* has been proposed by Richard Dawkins (1976) as an analogy to *gene* to describe “an idea, behavior, or style that spreads from person to person within a culture — often with the aim of conveying a particular phenomenon, theme, or meaning represented by the meme.” Memes are seen as “self-replicate, mutate, and respond to selective pressures” (Wikipedia *meme* page https://en.wikipedia.org/wiki/Meme). This similarity of genetic reproduction including mutation and the creative development of ideas has already been described by Humboldt (1794): *Not only the continuation of the species in the physical world is entrusted to the mutual procreation and reception. Even the purest and most spiritual sensation emerges in the same way, and even the thought, this finest and last offspring of sensuality, does not deny this origin. The mental procreative power is the genius.* My translation of “Diesem gegenseitigen Zeugen und Empfangen ist nicht bloß die Fortdauer der Gattungen in der Körperwelt anvertraut. Auch die reinste und geistige Empfindung geht auf demselben Wege hervor, und selbst der Gedanke, dieser feinste und letzte Sprössling der Sinnlichkeit, verläugnet diesen Ursprung nicht. Die geistige Zeugungskraft ist das Genie.”

perceive.

So it may be a special form of apophenia that we see causal relations where there are none. We are unsatisfied with, maybe even afraid of explanations that involve chance and existence out of nothing, without any 'causee'. Here the construct of a creator comes in handy, it gives the missing causal explanation of how it all came into being. It gives temporary satisfaction, if not digging too deep, as Russell explains: "*If everything must have a cause, then God must have a cause. If there can be anything without a cause, it may just as well be the world as God, so that there cannot be any validity in that argument. It is exactly of the same nature as the Hindu's view, that the world rested upon an elephant and the elephant rested upon a tortoise; and when they said, 'How about the tortoise?' the Indian said, 'Suppose we change the subject.'*"

We shall return to apophenia when talking about the interpretation of machine-generated texts in Chapter 6.

2.4 Sociobiology

The role of evolution in the Social Sciences is an ongoing debate. Under the form of the "nature versus nurture" debate it can be traced back at least to the 17th century debate between Locke and Descartes. The sociobiological view that all human sciences should strive to provide biological, i.e. evolutionary explanations to the observed behaviors was discredited by Social Darwinism, but it is currently seeing a quite spectacular comeback fueled by genetic analysis crossed with Big Data.

As an example of a biological explanation, consider a biological mystery that I have heard about only very recently: Sloths descend from their trees only for one reason: to defecate (Voirin 2013). It is clear that it is not a satisfactory explanation to state that they just feel like moving down from the tree; and even less acceptable is any reference to a sloth culture. The only admissible explanation is evolutionary: It somehow serves their chance of survival and reproduction to do so. This is why this is an open question in zoology: It seems utterly dangerous for a sloth to come down where it can fall prey to predators, and monkeys have no problem to do it from the canopy. The explanations of the sloth's behavior may be quite complex and including sexual selection, as exemplified by the peacock's tale that already Darwin invoked, but the current scientific paradigm obliges any explanations to be evolutionary by nature. Up to today applying the same logic to humans is commonly felt as disrespectful towards the human.

In linguistics, it seems that sociobiological explanations are mainly used for the explanation of the existence of language as such, for example in Hockett's (1955) language design features that include the theory of displacement, i.e. the idea that the capacity of talking about absent and abstract things has an evolutionary advantage. The arbitrariness of linguistic signs seems to restrict explanations from evolutionary linguistics to very global features of language, such as the influence of cognitive limitations to word length and dependency distance length (Menzerath xxxx, Gibson 1998, Liu 2008). Some evolutionary linguistics see language as a sexually selected feature just like the peacock's tail (Klasios, 2013). Together with the arbitrariness of linguistic signs, this could already make us abandon the idea that language has to be inherently simple. We will return to this point in Section 3.2 .

2.5 Is there beauty in simplicity?

"Simplicity is the ultimate sophistication." – Leonardo da Vinci

"Truth is ever to be found in simplicity, and not in the multiplicity and confusion of things." – Isaac Newton

"Everything should be made as simple as possible, but not simpler." – Albert Einstein

When the chaotic pantheon of multiple gods around the world had mostly been replaced by the story of the perfect monotheistic world, simplicity started to be an ideal of explanation. Saint Thomas Aquinas, mainly in his *Summa theologiae*, saw simplicity as the first and foremost quality of god (Weigel 2008) – he understood "simplicity" as "non-compositionality", *omnino simplex*. Of course, fitting the data to the monotheistic model caused some headaches to the theologians, for example when they asked questions of theodicy, i.e. how an omnipotent god could create evil. Equally, the creation of a unique and simple god did not come with a satisfying answer to the question of where god itself comes from as discussed in 2.3 . But still: The simplicity seems to have been a persuasive argument compared to polytheism.

We can only speculate why. Maybe a complex pantheon of gods took too much schooling to master. And as Russell (1927) convincingly argues, Christianity, the first missionizing monotheistic religion, was successful because it is resource efficient, in that it does not ask for any action from the side of the adherent. Faith was enough, which could be accomplished even by slaves.

Maybe there is also something more deeply inborn to the human mind's preference for simplicity. We know that many of our cognitive reasoning capacities are actually related to the application of social rules – and higher levels of cognition have possibly developed as a side effect of communal regulations. When trying to determine a culprit, it sounds reasonable that humans "naturally" prefer the simpler alibi as the better, possibly because this intuition encodes some primitive notion of probability. That might explain the preference of simpler explanations to more complex ones as a result of social adaptation (Pinker 1999, Chapter 5).

After the victory of monotheism, science aspired to take religion's place as the analytical and moral center of society. It was necessary to be just as attractive, just as awe-inspiring. The often-told story of Renaissance's (re-)discovery of the heliocentric universe posits simple and beautiful formulas that fit the observations even better than what was proposed by the holy texts. Bohr's atomic model looked similarly intuitive and Occam's razor gives the world a smooth and intelligible skin. The scientific myth states that simplicity is somehow built into the system: It is supposed that there exist simpler explanations that fit the data more closely. The awe lies in the simplicity of the system.

An example is electromagnetism which was the first "unified" theory of physics combining electricity and magnetism. With fewer equations and fewer hypotheses, it could predict more observations. This fueled the quest for more theories of this type. The expectation was that we would find something similar all over science, eventually possibly an overarching theory of all sciences.

Another interesting example from “hard” science that shows how nature is “ill-designed” and does not always come up with solutions of a simple and logical elegance: We often misunderstand Darwin in the sense that evolution finds a global optimum in its adaptation process. The aforementioned peacock’s tail is already a counterexample: It would be smarter for the whole species if peahen and peacock could agree to give up on the large tail (although they would then lose the advantage to being fostered by humans in parks). Even more astonishing is that photosynthesis itself is optimized for the pre-animal-life 2% oxygen level in the atmosphere, whereas the current atmosphere has a 20% level. Plants grow much better at a lower oxygen level (Moore 1981) but cannot redesign the enzymatic system from scratch to fit the current atmospheric conditions.

Whether the search for simplicity and logical elegance is justified in linguistics will be discussed in Section 3.3.2 .

3 On the history of linguistics

Science plays a double role today: Firstly, it is replacing religion in its role of paragon of virtue and central way of thinking, Secondly, it has the unpleasant task of lowering human self-esteem. This second point holds particularly true for linguistics, because Language is one of the central human hallmarks, while at the same time, in view of the shortcomings of our inborn behavior, it seems presumptuous to suppose that Language is anything but a specialized tool for some task – or it may possibly be the side effect of some useful feature. Language is not any more evolved than the elephant's trunk or the octopuses' color-changing skin. It even remains to be proven that high brain complexity makes us fit for survival. For the time being we are a very young species…

This finding goes together with the discovery of smart animal communication systems that are admittedly different from human communication but nonetheless efficient and often comparably complex, for example concerning entropy measures between human language and dolphin communication (McCowan et al. 1999). So one of the tasks of a contemporary linguist is to share the insight that Language is unsuitable as an awe-inspiring mysterious feature "surpassing material place". This is not an easy sell, because it lacks the feel-good factor of scientific results that show us how great and exceptional we are.

In its history, linguistics has been struggling with its place between natural and human sciences – choosing one direction over the other imposes specific paradigms on the study of Language, on how linguists work.

3.1 The "inner and universal perfection"[7] of Language

The role of the linguist's ancestor, the grammarian, was to give access to dead but sacred languages such as Sanskrit and Koranic Arabic, and the grammarian resisted for quite some time the temptations of seeing beauty in simplicity. The debate between grammarians who try to give a logical justification to grammatical rules and those that just list the rules seems to have been deeply entrenched when Ibn Mada[8] writes: *"And why is the agent in the nominative?" The correct*[9] *answer is […]: "This is how the Arabs speak"*, going against the quest of a justification of grammatical rules by overarching logical rules, and thus in favor of arbitrariness of the rules of an independent

7 Humboldt *On Language* p. 91. Page numbers of the English translation of 1988, used here in spite of the translation errors pointed out in Trabant (2012).

8 Ibn Mada was the first grammarian ever to use the term dependency in a grammatical sense comparable to the present day usage. He was born in 1119 in Cordoba, studied in Sevilla and Ceuta, and died 1195 in Sevilla. He is known for his only book, *Radd: the refutation of grammarians*, in which he tackles subjects that still seem modern today: He criticizes the use of ellipsis, i.e. underlying invisible forms, in the analyses of grammatical phenomena, for example when discussing whether the unmarked nominative form consists of a zero marker. I discovered this author in Graffi (2001) when researching on the origins of the notion of *dependency* in linguistics for the drafting of the introduction to the first International Conference on Dependency Linguistics (**Gerdes et al. 2011**). It is nonetheless controversial how far the proximity between the Arabic grammatical tradition and modern day grammar actually goes. Kouloughli (1999) points out that only relations were described where an actual morphological change is imposed by one word onto another. For this reason the Arabic syntactic structure did not include any invariable words such as complementizers and coordinating conjunctions.

9 This is certainly not an answer that we would see as "correct" today. It would be better to say that the nominative exists to make the semantic agent of the action identifiable.

syntactic module.

Similarly, Vaugelas (1647) was not yet guided by the search of simplicity while being aware that others look for reason in grammar when he writes: *In a word, Usage does many things by reason, many without reason, and many against reason. [...] Those are seriously mistaken, and they sin against the first language principle, who want to reason on our [language], and who condemn many generally accepted ways of speaking, because [these ways of speaking] were against reason; for reason is not to be considered at all.*[10]

His most famous quote, "*So here is how we define good Usage. It is the way of speaking of the healthiest part of the Court, in accordance with the way of writing of the healthiest part of the Authors of the time*"[11], stands in clear opposition to the "logical" rationale of grammatical rules that did not become dominant before the Port-Royal grammar and the creation of the Academie française, shortly after Vaugelas' quoted text at the beginning of the 17th century. The Academie française was founded with the explicit goal to "rendre [la langue françoise] la plus parfaite des modernes" 'make French the most perfect of modern languages'. It is not a surprise that Chomsky refers to the Port-Royal grammarians as the forerunners of Cartesian linguistics and Universal Grammar.

The Port-Royal grammar and the Academie française follow very different principles but share their appreciation of simplicity, as a descriptive or a prescriptive rule to follow. In contrast, Vaugelas' prescription was explicitly grounded in the putative aesthetics of the linguistic choices of the ruling class.[12] Thus, Vaugelas' word "usage" must not be understood as a statistical measure about how a language was spoken, and cannot be interpreted as the average usage of some social group such as the court at Versailles (cf. Wolf 1983 for further details).

When linguistics of the 19th century set out to include the newly discovered languages in their story, it was natural to discern an underlying logic below the structure of languages. The search for a family tree that resembles the aristocratic inheritance seemed natural in particular since it allowed to

10 The original text with modernized spelling: "*En un mot l'Usage fait beaucoup de choses par raison, beaucoup sans raison, et beaucoup contre raison. […] Ceux-là se trompent lourdement, et pêchent contre le premier principe des langues, qui veulent raisonner sur la nôtre, et qui condamnent beaucoup de façons de parler généralement reçues, parce qu'elles sont contre la raison ; car la raison n'y est point du tout considérée.*"

11 The original text: "*Voicy donc comme on definit le bon Usage. C'est la façon de parler de la plus saine partie de la Cour, conformément à la façon d'écrire de la plus saine partie des Auteurs du temps*"

12 Vaugelas is the same person who wrote in 1647 that "la forme masculine a prépondérance sur le féminin, parce que plus noble" 'the masculine form takes precedence over the feminine, because more noble', which is by the way the only prescriptive rule, invented by grammarians, that I am aware of, that actually made it into a language's grammar. In modern French, as soon as one masculine noun is present in a coordination, the agreement has to be masculine: "Les légumes et les fleurs sont **frais**" 'The vegetables and flowers are fresh' (*frais* 'fresh' being masculine as *légumes* 'vegetables', while *fleurs* 'flowers' is feminine). In old French and up to Racine the agreement was more commonly done with the linearly closest noun: "Ces trois jours et ces trois nuits **entières**" 'These whole three days and three nights' (*entières* 'whole' being feminine as *nuits* 'nights', although *jours* 'days' is masculine, Callamard 1998).
Prescriptive language has come back to the forefront of public debate with the introduction of "gender-inclusive language", which is sometimes prescribed by institutions or even countries (see for instance the Unesco Guidelines on gender-neutral language, Desprez-Bouanchaud 1999). For example, in languages with grammatical gender, a preference for epicene (common) nouns over gender-specific equivalents or the introduction of gender-neutral pronouns has repeatedly been shown to have the desired psychological effects on the audience of higher inclusiveness towards women (Sczesny et al. 2016). It is not clear yet whether and how this process will actually lead to a long-term grammatical modification of a language.

establish a hierarchy between languages. Family trees have unworthy bastards, so why not the world's languages? Humboldt (1836) sees "the lively and inseparable connection between languages and the mental capacity of nations" and has a whole chapter entitled "Less perfect language structures" (§23), starting his description with Semitic languages, Hebrew and Arabic, before moving to Delaware languages. Languages' extent of "cultivation" is measured by looking for "legacy from an earlier uncultivated phase" (p. 338). He proceeds by a "mere weighing of merits and defects" (p. 303) and places all languages on the two supposed extremes of Semitic languages on one side (something similar to what we would call *inflectional languages* today) and Chinese on the other side (akin to isolating languages); Sanskrit being the perfect language in the middle, which seems self-evident to Humboldt because its descendants, the Europeans, are the most successful.[13]

Reading Humboldt has something of a scientific (and emotional) roller coaster ride because of the incredible insights that Humboldt presents on the structure of languages, while at the same time fulfilling the duty of a propagandist of European expansionism, who has to prove a point about Western superiority, while presenting a universalist view of language[14] with some partly remorseful remarks on colonialism.[15] His constant value assessment of every language becomes utterly confused when Humboldt writes about Chinese (p.218): *"So although we gladly concede that the form of Chinese exhibits, more perhaps than any other language, the power of pure thought, and directs the mind more exclusively and urgently to this, precisely because it lops off all the small distracting sounds of connection, and although the reading of just a few Chinese texts reinforces this conviction to the point of admiration, still, even the most resolute defenders of this language can hardly maintain that it guides the mind's activity to the true center, from which poetry and philosophy, scientific research and eloquent discourse, spring forth with equal readiness."* Here, the text seems shiftlessly naive and the reader can feel Humboldt riven between the desire to prove the lower perfection of Chinese while facing the impossibility to deny sophistication to the Chinese culture.

Beyond the foolishness of some of his points and motivations, from a present-day perspective, the main observation to take away here is that for Humboldt the argument in favor of "cultivated" languages is their inherent logic, blending seamlessly in the tradition of Grammarians of the Academie française such as Girard (1747) who project logical reasoning into a language by means of the supposed "analogy" of some languages (such as French of course): *"Speech is therefore the most universal, natural, & graceful link in civil life; therefore the thing most deserving of being the object of its own exercises; especially in a polite & dominant nation, where the freedom to use*

13 "In the consideration of language as such, a form must be disclosed, which of all those imaginable coincides the most with the aims of language, and we must be able to judge the merits and defects of existing languages by the degree to which they approximate to this one form. [...] But the Sanscritic languages come closest to this form, and are likewise those in which the mental cultivation of mankind has evolved most happily in the longest sequence of advances. We can therefore regard them as a fixed point of comparison for all the rest." (p.216)

14 "Since the natural disposition to language is universal in man, and everyone must possess the key to the understanding of all languages, it follows automatically that the form of all languages must be essentially the same, and always achieve the universal purpose. The difference can lie only in the means, and only within the limits permitted by attainment of the goal." (p.215)

15 "The foreign settlers along the North American coastline pushed back the native inhabitants, and no doubt deprived them of their possessions, even in unjust ways; but they did not subdue these peoples..." (p.229)

[speech] was never constrained except by the rules of Reason."[16] In Humboldt's mind, too, proximity with mathematics and logic is the ultimate measurement of goodness. This proximity is supposed to materialize in some form of simplicity of the language ("the one form" of language) although Humboldt struggles to actually define this notion.

Humboldt was a discoverer and the world seemed infinite, magic, the human mind even more so. For him, "the brain has corridors, surpassing material place" as Emily Dickinson put it later. His famous "infinite use of finite means" (Humboldt 1836) is a theme that appeared much earlier in his thoughts, under the admiration of the variability of nature in general and is not at all restricted to language.[17]

When giving up gods as the ultimate cause of all being, Science did not give up on the mystery, the mystery of infinity. The 19th century scientist replaced the divine by the infinity of possibilities, fueled by the constant discovery of new things – and arrived at the wrong conclusion that the streak of discoveries can and will continue forever. So linguistics became a venture bound to prove language's infinity of possibilities and language's inherent logic and simplicity. Against all evidence (and some remaining voices, see Gross 1979) suggesting that languages are mainly idiosyncratic with frequent lexicalized rules and exceptions, and that languages are neither created to give freedom of thought nor to give access to logic. We will return to this point in Section 6.1.

3.2 The object of linguistics

The triumph of behaviorism in the Social Sciences in the first half of the 20th century also had an important influence on linguistics. Bloomfield's (1933) seminal work was grounded on behaviorist justifications, as required by the scientific fashion of his time, stimulus-response diagrams adorn his book (pp. 24, 25, 26, 31, 139), but few linguists actually bought into the behaviorist ideas as those thoughts seem superimposed on the structuralists descriptions. What the structuralist reader of the time took home is Bloomfield's language description and linguistic methodology, which are also more central to his work. Oddly enough, he is often seen as forerunner of phrase structure grammar. Yet, as we have explained in **Gerdes & Kahane (2011)**, Bloomfield's syntax is centered around the notion of *head* and the constituent is a derived, secondary object.

From a radical behaviorist point of view, exemplified by Skinner's Verbal Behavior (1957), Language is just a specific form of animal communication that should be studied using the same methodology. The individual languages are only interesting as far as they show how humans can be

16 My translation of *"La parole est donc le lien de la vie civil le plus universel, le plus naturel, & le plus gracieux ; par conséquence la chose qui mérite le plus d'être l'objet de ses propres exercices; surtout chez une nation polie & dominante, où la liberté d'en faire usage ne fut jamais contrainte que par les règles de la Raison."* (Girard 1747)

17 Many years before turning to language studies, Wilhelm von Humboldt wrote about sexuality, art and genius, about what we would call "creativity" today in his *On Sexual Difference* 1794: *Nature, which pursues infinite purposes by finite means, bases its building on the opposition of forces. Everything that is limited aims at destruction, and heavenly peace dwells only in the sphere of what is self-sufficient. The destructive activity of the one must therefore be opposed by the other, and by thwarting each other's final purpose, they fulfill Nature's unrestrained plan.* My translation of "Die Natur welche mit endlichen Mitteln unendliche Zwecke verfolgt gründet ihr Gebäude auf den Widerstreit der Kräfte. Alles Beschränkte zielt auf Zerstörung und der himmlische Friede wohnt allein in dem Wirkungskreis dessen was sich selbst genügt. Der zerstörenden Thätigkeit des einen muss daher das andre entgegenstreben und indem beide gegenseitig einander ihren Endzweck vereiteln erfüllen sie den schrankenlosen Plan der Natur."

trained to use different languages. Hockett (1955) explicitly sees himself in a Bloomfieldian tradition when he presents “a mechanical and mathematical model of a human being regarded as a talking animal” heavily influenced by notions from Shannon’s Information Theory. Hockett’s syntactic module, named Grammatic Headquarters (G.H.Q.) in a military post-war fashion, is modeled as probabilistic transitions between morphemes, anticipating n-gram parsing by decades. Illustration 1 shows the center of interest of behaviorist linguistics.

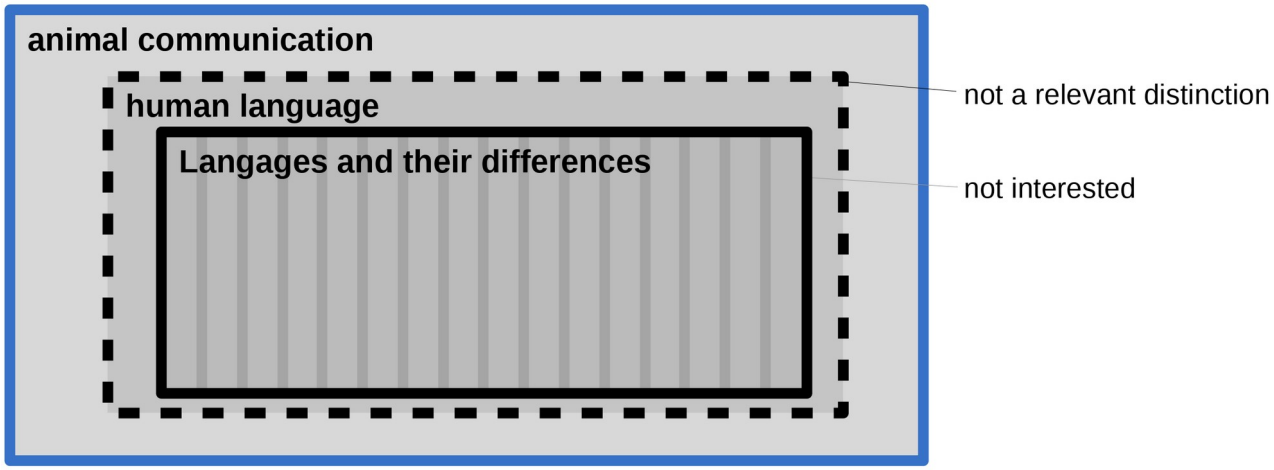

Illustration 1: Radical behaviorist linguistics, primary object of study in blue. Interest in individual languages and their differences is seen as a subset of interest in human language in general.

Strangely enough, in spite of the apparent modernity from a Machine Learning point of view, Hockett’s influence and “fame” on linguistics consisted mainly in serving as the counter-example of what should be done in linguistics, as he and Skinner serve as the target of Chomsky’s criticism of all empirical linguistics (Radick 2016).

Chomsky’s refusal of empirical methods of behaviorism was the result of a gradual change in its early work. He only started repudiating behaviorism when he noticed how it contradicts his views on Language, but grew more and more radical over the years, all the way to being “comfortable with the notion that language had not gradually evolved at all, but, in accordance with a still-undiscovered physical law, had emerged, in one go, once the brain had evolved beyond a threshold of complexity reached only with the human species.” (Radick 2016)[18] This gives a near-divine status to Language and is reminiscent of the current day “Intelligent Design” attempt to curtail Darwinist biology in US schools. Later, Chomsky (1971) goes as far as stating that “what defines behaviorism is the very curious and self-destructive assumption that you are not permitted to create an interesting theory. […] I don’t think that behaviorism is a science” comparing its obsession with

18 Chomsky (1968) states explicitly “When we ask what human language is, we find no striking similarity to animal communication systems. There is nothing useful to be said about behavior or thought at the level of abstraction at which animal and human communication fall together. The examples of animal communication that have been examined to date do share many of the properties of human gestural systems, and it might be reasonable to explore the possibility of direct connection in this case. But human language, it appears, is based on entirely different principles. This, I think, is an important point, often overlooked by those who approach human language as a natural, biological phenomenon[;] in particular, it seems rather pointless, for these reasons, to speculate about the evolution of human language from simpler systems—perhaps as absurd as it would be to speculate about the “evolution” of atoms from clouds of elementary particles.”

data to trying to see physics as a "meter-reading science". The word "interesting" is central here: A data-driven approach may be more empirical, fitting the data better, but it will not be interesting. "Interesting" is thus eponymous to some underlying universal simplicity of Language whose existence is presupposed.

So we can distinguish a biological Darwinian approach to Language, that wants to explain language as one case of animal communication, and Generative Grammar refusing this connection to evolution, and explicitly positioning linguistics as a subfield of psychology. Illustration 2 provides an overview of the focus of Generative Grammar.

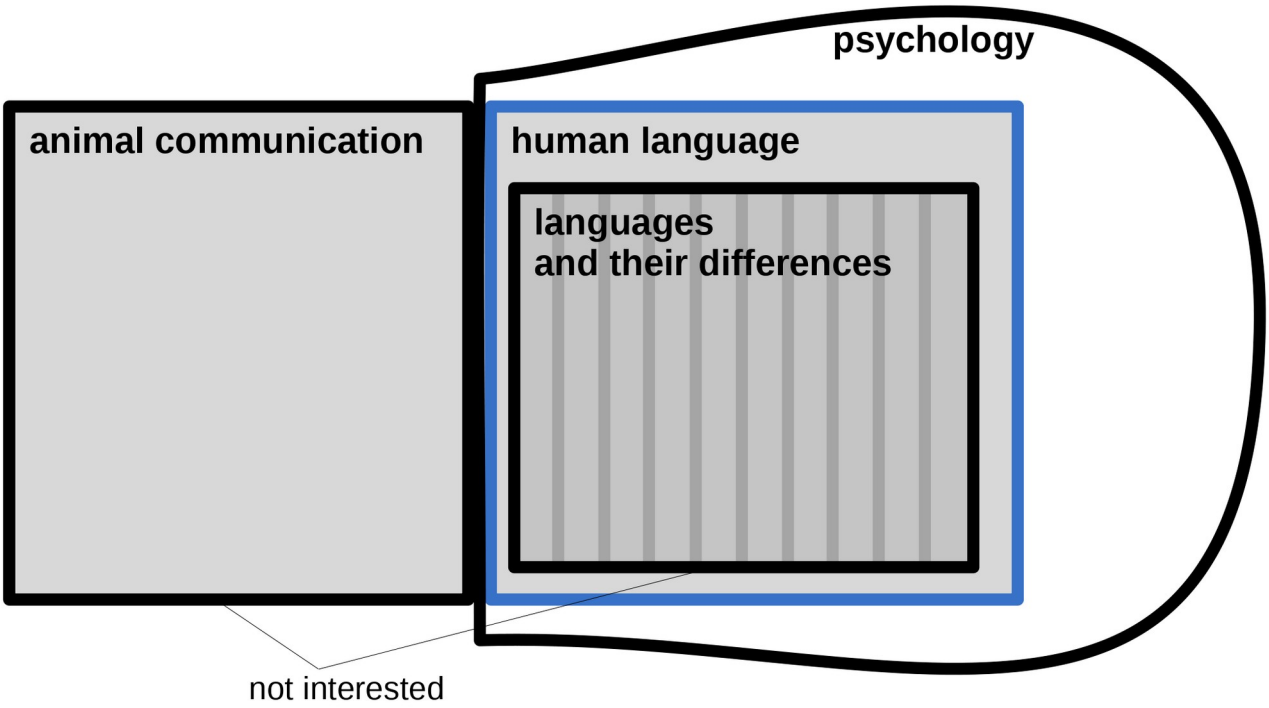

Illustration 2: Generative linguistics, primary object of study in blue.

It is noteworthy, and has often been criticized (see for example Golumbia 2004 on the biased notion of "non-configurational" languages), that the Universal Grammar rules that need to be discovered, principles and parameters alike, only require English. Yet, English is just one instance of Language. Non-English speakers are seen as doing astonishing language acrobatics – to think English and speak so weirdly differently. Haider (1993) criticizes the Chomskyan model for its description of German, observing that he fails "to understand why English should better reflect Universal Grammar than German in any interesting sense."[19]

19 Complete context: *"At the beginning of this section I want to put my credo, borrowed from Albert Einstein in a modified form: The grammar is subtle, but not malicious. However, the opposite impression arises immediately when one considers the contortions compelled by an analysis of German following directly the Anglo-Romance standard model forces. If I am faced with the choice of deciding whether the grammar of German seems odd because German is odd, or because we apply the theory oddly, I will stick to the safe option, i.e. the latter. I fail to understand why English should better reflect Universal Grammar than German in any interesting sense."* Haider 1993:142, quoted by Kathol 1995:1 Original text: *"An den Anfang dieses Abschnitts will ich mein in abgewandelter Form Albert Einstein entlehntes Credo stellen: Die Grammatik ist raffiniert, aber nicht boshaft. Der gegenteilige*

This situation can be opposed to the mainly Central and East European continuation of structural syntax, primarily around the Prague school and Mel'čuk's Meaning-Text Theory. Here the focus is on an implementable description of different languages. These approaches share the unordered tree of words, the dependency tree introduced by Tesnière, a slavist (1959), as the central representation of syntactic structure[20], and we can refer to them as structural dependency grammar. Just like Generativism, the language rules are supposed to be discrete and simple, and corpora or other forms of frequency information are not part of the methodology – corpora in the sense of large collection of coherent spontaneous language data, as opposed to lists of examples, often quite large, that are used since Damourette & Pichon (1911-1944). One important difference lies in the focus on idiosyncrasies: on language differences as well as on exceptions from simple rules, and thus on the integration of the dictionary in the model of syntax. Illustration 3 gives an overview of Structural Dependency Grammar's centers of interest.

Another difference, to be discussed in Chapter 6, and eponym of Mel'čuk's Meaning-Text Theory, is the integration of semantics in the model. Moreover, structural dependency grammar has no problem with the integration of human language as one form of animal communication; if other forms of communication are close enough to human language, they could be modeled with a similar formal machinery. Yet, dependency linguists have an approach not unlike many typologists' techniques as they see themselves as specialists of their own language, their mother tongue, or the language they specialize in, busy with the description of the language's countless idiomatic structures. They leave the study of general constraints that govern language and possibly other non-linguistic cognitive features to researchers in biology or cognitive science. Universal Grammar may be the end result, not an *a priori*, of the analysis: Once all languages will be described, we might find common rules between all of them. It will then be up to research to determine whether these common rules are due to cognitive constraints, common origin of all languages, chance, or other factors yet to be discovered.

Eindruck stellt sich aber unmittelbar ein, wenn man betrachtet zu welchen Verrenkungen eine dem anglo-romanischen Standardmodell direkt folgende Analyse des Deutschen zwingt. Stehe ich vor der Wahl, zu entscheiden, ob die Grammatik des Deutschen verquer scheint, weil Deutsch verquer ist, oder weil wir die Theorie verquer anwenden, so halte ich mich an die sichere Möglichkeit, nämlich die letztgenannte. Mir will es nicht einleuchten, dass Englisch in irgendeinem interessanten Sinn die UG besser reflektieren soll als Deutsch."

20 Kahane (to appear) shows that a precise description of the relations that constitute a dependency analysis was already present in much earlier works such as Buffier 1709 although the dependency tree was not yet drawn, as the whole idea of a scientific diagram appeared only in the first half of the 19th century. Note that Buffier was a Frenchman born in Poland, which is in line with my (possibly anecdotal) claim that language knowledge influences linguistic theory.

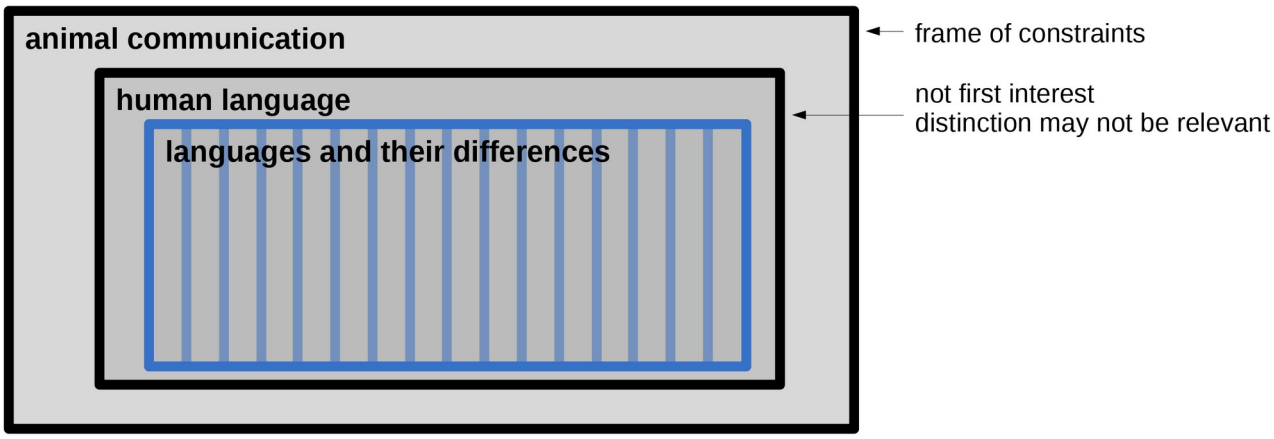

Illustration 3: Structural dependency grammar, primary object of study in blue
In this view, human language is just one type of animal communication, and it is not supposed a priori *that a significant border exists between animal and human communication*

3.3 Non-empirical Linguistics

So the subterfuge that Chomskyan grammar brought to perfection was twofold when Generative Grammar declared that:

- Language must have simple underlying rules that can create an infinity of sentences with a few simple rules.
- We actually want to describe an ideal language, called the Language Competence, which is supposed to live in the neural system of the native speaker. The counter-examples to simple rules that are staring right in the face of every syntactician are akin to measurement errors and called “performance errors”.

The simplicity of the rules makes sense if we have good reasons to expect that the underlying rules are actually simple. This point was also backed up by an “empirical” observation that proved to be wrong: the poverty of stimulus. Simplified, the idea is that children do not get enough language input to be able to deduce the whole grammar of the language. It follows that an inborn Universal Grammar exists as a biological feature shared among all human languages, and the language input that children have access to is only instantiating the existing rules. Yet, the argument has methodological flaws (Pullum & Scholz 2002) and modern machine learning systems have already been able for a few years to deduce pretty good rules and generalizations on much less language data than what is available to the newborn – and without nearly any preconceived rules (Clark & Lappin 2010). Yet, we cannot conclude that languages are arbitrary. Of course, cognitive limitations of the brain and biological evolution of the relevant brain areas will shape human language. The Magical Number Seven (Miller 1956) relates working memory capacity with phonetic and syntactic chunking. In a joint work with Cédric Gendrot and Martine Adda-Decker we showed that for French journalistic text, syntactic chunking that tries to build segments of length 7 actually coincides best with actual prosodic boundaries (**Gendrot et al. 2009, 2010, 2011**). While we are currently witnessing a revival of cognitive approaches to linguistics (Croft & Cruse 2004), “the link

between computer processing and cognition has become very tenuous if not completely lost"[21] (Poibeau 2011) – as very little research in Natural Language Processing claims cognitive viability of their systems. Anyhow, the study of cognitive limitations may not actually lie in the domain of linguistics as they affect all parts of cognition – a point vigorously defended by Igor Mel'čuk (personal communication).

3.3.1 Infinite competence

Note also that the opposition of Competence and Performance is much older than Chomsky. It is a natural result of the epistemological context in which modern structuralist linguistics evolved, as described in Section 3.1 . The opposition has existed in similar forms since at least Humboldt (see Bußmann and Lauffer 2008): Ergon – Energeia (Wilhelm von Humboldt 1836), Sprache – Rede (Hermann Paul 1880), Sprachsystem – aktualisierte Rede (Georg von der Gabelentz 1891), Langue – Parole (Ferdinand de Saussure 1896), Sprachgebilde – Sprechakt (Karl Bühler 1934), register – use (Michael A. K. Halliday 1961), Competence – performance (Noam Chomsky 1965)[22] and the hypothesis of the existence of this opposition does not presuppose the idea that linguistics should be interested in simplified rules of an ideal language.

The limitation to the "core" of Language, to the regularities that are governed by parameter settings, has been identified as central to the development of structural linguistics. It is only at that "core" that exact regularities can be discovered (Hajičová 2011). The study of the "periphery", the exceptions, is put off to a later date. The problem is of methodological nature, because the distinction between core and periphery (or between competence and performance) is defined on the fly, dynamically while describing a language. This allows identifying regularities nearly *ad libitum* because it allows excluding all contrary data points that one might encounter to the periphery.

The construction of Competence as linguistics' only subject matter remains the main point of Chomskyan linguistics as it discredits the work on language usage as a subaltern interest in experimental setups and measurement errors, and it accredits the refusal of empirical methods. This refusal of corpora as the adequate empirical basis of linguistics is a hypothesis that seems hard to maintain.[23] Again the supposed infinity of Language was central in the argument and it appears in

21 My translation of "Le lien entre traitement informatique et cognition est devenu très ténu quand il n'a pas purement et simplement disparu" (p. 12)

22 Although a latecomer, Competence vs. Performance has become today's standard terminology probably mainly for reasons of better public relations.

23 My personal encounter (or should I say crash?) with Generative Grammar took place at the 7th Germanic Linguistics Annual Conference (GLAC'7) in Banff, Canada (**Gerdes & Kahane 2001**). I presented a set of word order variations for the same dependency tree that I had computed from my topological model (see Section 4.2.1). When the slides showed the sentences, my presentation was disturbed by people talking loudly and some voices shouted "word garbage". The session chair had to intervene and ask people to sit down again and let me finish. Although I had not expected such an uproar, I was prepared for a mined territory, and for every single word order variation that I presented, I had found a real sentence in some newspaper with the same structure. What followed after my presentation was even more astonishing than the interruptions of my talk: Among the many people that came up to the front was a Canadian graduate student of Germanic syntax. At some point he said in front of everybody who was busy pointing out the errors of my presentation: "I had been working on German for many years without actually understanding the word order. And I have just finally got it thanks to this talk." This silenced the ongoing conversation. What is more is that Werner Abraham, arguably the most important Generative Grammarian of Germanic languages, came up to me later and took a lot of his time to explain why my attested word-order examples are performance errors, if viewed from the larger picture of Universal Grammar. I have

the very first sentences of Chomsky's Syntactic Structures (1957): "*From now on I will consider a language to be a set (finite or infinite) of sentences, each finite in length and constructed out of a finite set of elements. All natural languages in their spoken or written form are languages in this sense, since each natural language has a finite number of phonemes (or letters in its alphabet) and each sentence is representable as a finite sequence of these phonemes (or letters), though there are infinitely many sentences. Similarly the set of 'sentences' of some formalized system of mathematics can be considered a language.*"

Language is defined as infinite, which stands in contrast to the finiteness of corpora. In the refusal of corpora as the adequate empirical basis of linguistics, Generative Grammar preserved linguistics from the arrival of probabilities much longer than other Social Sciences. But maybe there was a further reaching intuition behind this that allowing empirical data, i.e. corpora, into linguistics would kill structural linguistics itself. As long as information technology only handled very small corpora, Chomsky's points against corpus linguistics (as being skewed and inadequate) remained at least partly valid and armchair linguistics the methodology of choice (Fillmore 1988) in spite of the obvious problems of the measurement of grammaticality, as a binary and individual notion.

While not wishing to descend into technological determinism, it has become impossible since the emergence of powerful computational systems within the linguist's reach to ignore actual linguistic data. However, the hope that Big Data would finally give us access to the simple rules underlying Language has not been fulfilled. It was much worse: Big Data has shown us not only that the simple language rules do not exist, putting into question the linguist's quest for the last centuries. Machine Learning seems to indicate that rules exist, but they are not intellectually satisfying in any way because of their sheer complexity. We do not want to take these rules as an answer. The situation is a bit like in Douglas Adams' "Hitchiker's Guide to the Galaxy" where the computer gives "42" as the long-awaited answer, forcing everyone to wonder what the question had actually been. Or, to give an actual historic example, linguistics, just as much of the rest of the Humanities, has to take a blow comparable to the one that shook Physics when quantum approaches gave an unexpected but most importantly counter-intuitive answer to long-standing questions by replacing absolute rules by probabilistic equations.

3.3.2 Simple rules

We understand now how our cognitive limitations and social and personal desires have misled Science to ask the wrong questions and to accept the wrong answers.

We prefer a language description with 10 categories to a language with 1000 categories, or a grammar with 100 rules to a grammar with a million rules – not because there is an evident intrinsic value of a shorter rule set over a longer one to empirically fit the data better but rather because it fits our cognitive limitations better. Popper already "discards the instrumentalist solution that a theory is better than its rival if it is 'simpler' than its rival from the point of view of intellectual comfort" (Lakatos 1968).

benefited a lot from these explanations, and Werner Abraham later even agreed to be a member of my PhD jury. Ultimately, this dialogue has led me to an in-depth criticism of Generative Grammar's syntax representation in **Gerdes (2006)**.

Of course, if the 100-rules grammar would be exactly as empirically grounded as the million-rules grammar, the million-rules grammar would simply contain redundant information. – It would either be repetitive or not have captured some sort of generalization. I am not arguing that Occam's razor is wrong: Two descriptions with the exact identical empirical basis might have different lengths. Then the shorter version can be considered as more beautiful, more pleasing to the scientist's eye. But maybe the different size of the two descriptions may just not matter as much as we had thought. Classical Science put simplicity on a pedestal where it only belongs if we want our theory to be understandable by a single human that does not use intelligent tools.

In the Humanities in particular, we may hold the false belief that Occam's razor promises us the existence of a simple rule. The simple overarching rule may exist in Physics – although even there we observe the emergence of a plethora of elementary particles – it becomes increasingly clear that simple rules are a lousy fit for Language. Language may not simply have many exceptions. It may be nothing but a set of idiosyncratic rules and any general rule may be abusive, see Section 4.2.2 for an extensive example.

Modeling Language on the ideal of Physics is surprisingly long-lived. The exceptions to every single rule that linguists put forward did not discourage structural linguists from keeping it up. After all, Physics also has measurement errors[24] – the underlying theory only shows up statistically, as a mean of measures. Moreover, some models are preserved although they are known to be incomplete. An example is the Bohr model of atomic structure, which "may be considered to be an obsolete scientific theory. However, because of its simplicity, and its correct results for selected systems [...], the Bohr model is still commonly taught to introduce students to quantum mechanics or energy level diagrams before moving on to the more accurate, but more complex, valence shell atom" (Wikipedia page of the Bohr model[25]). More generally, "scientists [...] tend to ignore counterexamples, or as they prefer to call them, 'recalcitrant' or 'residual' instances; and elaborate and apply their theories regardless. Their theories frequently show remarkable tenacity" (Lakatos 1968). This is what Kuhn (1962) denounces as dogmatic "normal" science.

In linguistics, too, when a Popperian falsifier appears, such as a counterexample to a grammatical rule, we cannot directly conclude that our theory must be rejected. The counter-example refutes the grammatical rule, but it may indeed be a false observation, a misinterpretation of the data. Of course, "experiments do not overthrow theories, as Popper has it, but only increase the problem-fever of the body of science. […] There must be no elimination without the acceptance of a better theory" (Lakatos 1968). What is special about linguistics is firstly that Generative assumptions prevailed in spite of the existence of a better, simpler, more empirically correct theory, Dependency Grammar. Note though that Dependency Grammar makes reduced claims about Language as a whole, as I have shown in Section 3.2 , which makes Dependency Grammar's results of virtually

24 When studying Mathematics in Berlin, I saw that physics is the source and motivations of many mathematical structures and theories. So, attending Physics 101 at the next-door institute seemed the thing to do not only because making metal balls roll down slopes was a nice change from Mathematics. It came as a surprise how many of the actual measures did not fit the theory and had to be excluded. I happily returned to my Mathematics department where I was not forced into what I felt was intellectual treachery, where I could think about flawless objects such as sets, groups, and rings. Perfection.

25 https://en.wikipedia.org/wiki/Bohr_model, consulted in July 2018.

no interest to a Generative Grammarian. What is important here is that all of structural linguistics presumes the existence of simple grammatical rules that need to be unveiled (see also Section 4.1.1). Yet, there is no good reason to postulate their existence, and at the same time there are many good psychological motivations that make us want to believe in them.

3.4 Linguistic innumeracy

It would be ahistorical to attribute the success of non-empirical linguistics to Chomsky's groundbreaking ideas and convincing personality alone. Again, he was nothing more, but also nothing less than the messiah who fit the needs of the linguistic community of his era.

His work must be seen in the context of his time: *Syntactic Structures* was published in the year of the Sputnik Crisis and the beginning of the Space Race. The Russians thus clearly had knowledge that needed to be understood. So Machine Translation was an urgently needed tool beyond the usual spying in the Cold War. The increasingly evident failure of pure computational methods to achieve usable machine translation made it clear that linguistic knowledge needs to be developed and that the equation of deciphering Russian language with the successful deciphering of the German military message encoding was misleading. In this context, Categorial Grammars were proposed by Bar-Hillel (1953), who worked at MIT and who was responsible for the US Machine Translation program. And Chomsky had also been working at MIT since 1955, a university that up to today is principally financed by the Pentagon.

At the same time, the roaring fifties and Johnson's Great Society in the sixties brought an enormous increase of students to the universities, also women and minorities, with the promise of social mobility thanks to education. In this context, Social Sciences increased more than hard sciences. Many people found jobs in universities that would not be hired in less revolutionary times. Chomsky was thus in a very privileged position to disseminate his new theories, actually any theory. The development of linguistics can be seen in the context of Social Sciences resumed by Payne (2014): "*The 1960s brought in symbolic interactionism, phenomenology, ethnomethodology, and social constructionism, while the next decade saw Marxian disputation supplanted by feminist sociology. This was followed by post-modernism in which discourse shifted to the cultural contextualisation of words and ideas.*" This led to an "epistemological anomie" without "*any obvious criteria for adequately certifying knowledge, nor for judging the appropriateness of methodological procedures, nor even for what are suitable topics for investigation*" (Bell & Newby 1977). And in these disorienting times, at least linguists were lucky to have a clear goal: describing Universal Grammar.

Chomsky's theories were explicitly qualitative, already ridiculing in his *Syntactic Structures* any quantification, which was a logical consequence of the behaviorist experimental setup put forward by Skinner and Hockett. Chomsky goes as far as explicitly excluding the interpretation of his grammar as a simplification of statistical empirical data. "*It is natural … to assume that the linguist's sharp distinction between grammatical and ungrammatical is motivated by a feeling that since the "reality" of language is too complex to be described completely, he must content himself with a schematized version replacing "zero probability, and all extremely low probabilities, by*

impossible, and all higher probabilities by possible." We see, however, that this idea is quite incorrect, and that a structural analysis cannot be understood as a schematic summary developed by sharpening the blurred edges in the full statistical picture." Chomsky thus excludes the possibility of interpreting competence simply as a threshold of decipherability and frequency of a sentence. His linguistics did not need statistics.

An epistemological explanation could attribute the success of qualitative analysis and discrete values to the "lack of numeracy" in all the Social Sciences, where statistical classes where often outsourced to starting mathematics teachers with no knowledge of the domain of the students (Payne 2014, describing the development in the United Kingdom, but we can suppose the situation to be similar in other Western countries). This is not a recent problem as von Gabelentz already noted in 1891 that "*it hardly needs to be mentioned that our science [linguistics], like any other, requires a mind that is trained in logic. However, there is a difference between the theoretical preoccupation with logic and its practical application. It is well known that at many universities the lectures on this science [of logic] are among those which are more enrolled into than actually attended.*"[26] We are facing here a problem of self-reproductivity, because up to today, no curriculum in linguistics that I am aware of would prepare students to teach even an introductory class of mathematical methodology such as Statistics.

Although linguistics established its methodology influenced by other Social Sciences, it remained heavily impressed by natural sciences as their declared ideal, more specifically by physics and mathematics. Biology, on the contrary, was not looked up to although or possibly because of the possible natural subordination of linguistics to biology. Biology was the infamous "stamp collection", see Section 5.4 , compared to physics, the "real science". The adoration of physics and mathematics in combination with the widespread innumeracy and ignorance of empirical methods among social scientists may also have led to what Bricmont and Sokal have called "Intellectual Impostures" (1998).

Again, in spite of the fear of falling into scientific determinism, we would like to ask: What other way than Generative Grammar could have been taken by a scholar of linguistics, who (a) wants to find beautiful simple rules à la $e=mc^2$, (b) has no training in statistics, (c) only masters English[27], and (d) has a slight uneasiness with the idea that humans are not higher developed than other animals? Put this way, the generative path that linguistics took seems pre-determined. Equally, if I were a linguist who (a) wants to model discretely the language translation process, and (b) knows Slavic languages (and possibly Ancient Greek, Latin, German, and Romance languages as Tesnière), it would seem impossible to get stuck with rewriting grammars and Universal Grammar. It seems likely that I would head straight to unordered dependency trees and to a semantic representation that looks a bit like these trees.

26 My translation of "*Dass unsere Wissenschaft wie jede andere einen logisch geübten Geist voraussetzt, braucht kaum erst erwähnt zu werden. Es ist aber doch ein Unterschied zwischen der theoretischen Beschäftigung mit der Logik und ihrer praktischen Verwerthung. Bekanntlich gehören an vielen Hochschulen die Vorlesungen über diese Wissenschaft zu denen, welche mehr belegt als besucht werden*"

27 This does not fully apply to Chomsky who has at least a good theoretical knowledge of Hebrew, with both his parents being specialists of that language and his B.A. thesis describing the "Morphophonemics of Modern Hebrew" (Barsky 1997).

And do we have any other choice today than doing science with neural networks?

4 On formal grammars

Formal grammars are the mathematical branch of structural linguistics. It tries to use mathematical descriptions and reasoning to answer linguistic questions and it can thus be seen as one type of applied mathematics. Naturally, the field is mainly composed of researchers who have at least some part of mathematical training. It has taken off in the 1980s, when Generative Grammar was at the pinnacle of its influence, and it fits well into the surrounding paradigm of its time as it can be conducted nearly without recurring to empirical data. In this chapter, I will describe my path through formal grammars and show how this field can give very satisfying answers to long-standing questions of syntax, but at the same time, how it constantly runs into the limits of Language's complexity.

4.1 Example of Tree Adjoing Grammars and Metagrammars

The fate of Tree Adjoining Grammars (TAG, Joshi 1985) is a perfect example for the wild waters that formal linguistics got into in times of Big Data and powerful machines. For mathematicians, TAG were a natural entry point to linguistics, also for me. I had just finished my Master thesis in Arithmetic Geometry, (**Gerdes & Kempe 1996**), describing an algorithm that makes the decomposition of algebraic varieties into equi-dimensional parts polynomial[28]. And looking at a syntactic formalism that could be analyzed in polynomial time was the most natural of things (**Gerdes 1998**).

4.1.1 What makes TAG so attractive?

Let us temporarily accept, for the sake of the argument, that we want it all: We want to understand Syntax, i.e. the submodule of the human language capacities that combines words to sentences.[29] We are ready to believe that syntax is something like an autonomous system of the cognitive system as brain studies suggest – because it can break down more or less independently of other cognitive functions. We also know from psycholinguistic studies that long sentences do not explode our heads, and we can understand language in real time. So we look for a model with a "plausible parsing model" – a description that takes exponential time when combining words to sentences, such as Transformation Grammars, is cognitively implausible.[30] There are of course a few caveats to the direct application of the algorithmic parsing properties to the actual evaluation of language models, one being the fact that many parsing algorithms can be highly efficient in reality and the

28 An algebraic variety is the set of solutions of a system of polynomial equations. A single two-dimensional polynomial equation such as $xy=0$ is true if either x or y is equal to zero. The set of solutions is thus the x and the y axes themselves, it forms a + sign. Equally, $y(y–1)=0$ holds if either $y=0$ or $y=1$, thus forming two horizontal lines in affine space, akin to a = sign. If we now combine these two polynomial equations into one system $\{xy=0, y(y–1)=0\}$, these equations are true in the intersection of the two zero-loci which looks like ÷, a point and a line, i.e. a zero-dimensional object {x=0, y=1} and a one-dimensional object {y=0}. Finding the algebraic descriptions of these two equi-dimensional objects (the point and the line) is algorithmically complex in a general setting, but can be useful for the interpretation of measurements such as radar distance data.

29 This is not really an operational definition as we would need to be able to define *words* as well as *sentences*, both of which is rather a consequence than an *a priori* object of syntax, see **Kahane and Gerdes (in preparation)**.

30 Note that these presuppositions have a strong 1990 feeling to them, because even better computational properties, would not make the model cognitively plausible.

algorithmic measurement of the worst-case scenario is an asymptotic property that does not show in actual psycholinguistic analyses (cf. for example Berwick and Weinberg 1982). It may even be impossible to construct a whole grammatical sentence in any language whose parse grows exponentially by adding words. Another problem, to be discussed in Section 5.7 , is that parsing, i.e. giving structure(s) to a sentence in order to decide whether the sentence is grammatical, is very different from actual language understanding, i.e. linking sentences to meaning.

And we have faith in Occam, and thus we believe that simpler answers are better answers. For syntax, this boils down to the belief that somewhere an easy explanation of syntax is hiding: an aesthetic model that pleases our weak brain with simple symmetries, a model that guides how the neurons of the relevant parts of the brain are connecting. Put simply, a few simple rules are only waiting to be unearthed that would be empirically accurate and computationally fast.

It was easy to show that simple rewriting rules do not fulfill the requirements. They do not capture in a satisfying manner the observed variations in language, in particular long-distance dependencies. Also, each erratic behavior of some lexeme requires a special feature that is present in general rules, independently of the presence or the absence of the lexeme. For example, we would need a special feature for each set of exclusions such as verbs that do not allow passives (**5 Euros was costed by the book*) or some adjectives that allow a so-called 'tough construction' only with an adverb of degree such as *too* (*an easy to read book* vs. **a big to read book*, but *a book too big to read*).

Here, Tree Adjoining Grammars (TAG) enter the scene. They can be analyzed in polynomial time (which sounded like an interesting feature in the 1990s), and they are lexicalized: Each word form comes with a set of trees that indicate how the form can connect with other forms. They allow for a satisfying analysis of some syntactic phenomena that rewriting grammars were struggling to tame. They may not perfectly fit to the data either, for example the problem of recursively embedded verbs with their arguments in Dutch (Rambow & Joshi 1994). But this can be fixed by some add-ons to the system such as multi-component TAGs (Weir 1988) and anyhow even simple TAGs can still describe the data up to a certain degree of embedding (three verbs). This can be sold as a feature, not a bug, as humans also struggle with the so-called center embeddings (Joshi et al. 1994).[31]

There is another feature of TAG that is noteworthy and which keeps some people working on TAG up to today: An analysis with a Tree Adjoining Grammar always creates two trees in parallel: A phrase structure tree that tries to resemble a surface tree of Transformational Grammars, called *derived tree*, and a hierarchy of the sentence's lexical units, called the *derivation tree*, that resembles a dependency tree. TAG thus constitutes a gentle path for mathematicians to move from phrase structure trees to dependency trees. It is a way to abstract away from word-order constraints of the language, because dependency trees are unordered and deeper (closer to semantics) (Rambow & Joshi 1994, Candito & Kahane 1998).[32]

31 Center embeddings are sentences where the additional information is added in the middle of the base structure, thus distancing the elements of the base structure. An example with two center embeddings shows already the difficulty to understand could be "The guy to whom the lady who just sneezed talked all night came to see me." which can be understood much more easily when taking the sentence apart in this paraphrase: "The guy came to see me to whom the lady talked all night – the lady who just sneezed."

32 Note that Lexical Functional Grammar (LFG, Kaplan & Bresnan 1982) can be seen as another topological

Inversely, TAG could thrive in the US in spite of the unrelenting dominance of Generative Grammar because firstly it was mainly practiced in computer science departments, but also because the shape of the actual tree-segments does not matter too much. As soon as the attachment nodes are present, the tree-segment is complete, and it is possible to add a few nodes nearly free of cost, in particular those required by generativists, such as IP nodes (inflectional phrase, fashionable since the 1980's). Although this emptied the generativists' phrasal nodes of their meaning, which is the linear delimitation and head-determination of phrases, it nonetheless made the resulting phrase tree look reasonable to generative grammarians.[33]

4.1.2 The limits of TAG

Not only verb embeddings are difficult for TAG. The formalism also hits its limits with another construction: extraposed relative clauses. In German, the standard position for a relative clause is the so-called Nachfeld (post-field), behind all the main constructions of the sentence. On this subject I had my first methodological debate with my PhD adviser: Could that astonishing standard position of relatives mean that in German, the noun phrase (NP) and relative clauses do not actually form a unit and that the relative pronoun is nothing but a standard referential pronoun? The main argument against this analysis is that the relative clause cannot appear without preceding NP and cannot refer to NPs beyond the sentence boundary. But the temptation was strong to actually recode the grammar in order to make German relatives fit into the mathematically beautiful polynomial formalism of Tree Adjoining Grammars. If you refuse to re-analyze German in accordance to TAG, the only way out is to trick a little: German extraposed relatives can be described in a TAG grammar by pushing lexical information into agreement features that are unified across other tree-segments belonging to words that have nothing to do with the actual NP and its relative clause. This is a violation of TAG's so-called "strong co-occurrence constraint" (**Gerdes 2002a**). But hey, it makes German fit into TAG!

Other difficulties stemmed from the very specific German word order discussed in Section 4.2.1 . But this, too, could be solved by adding a few specific features to the trees that checked for the order restrictions. The hardest blow to TAG seems at first to be of practical nature: The number of trees. As mentioned above, TAG attaches to each word form the partial trees that anticipate the connections the form can engage in. That does indeed allow adding and removing those tree segments from specific lexemes or to add restrictions on its arguments. But even without going that much into detail, without encoding any lexical idiosyncrasies such as selectional restrictions, the simple crossing of valency and word order produces a large number of possible combination for each word form. In order to automatically compute the complete paradigm of combinations for each word form, it has been proposed to describe the possible combinations in an inheritance lattice of properties, called a meta-grammar (Candito 1999, Gaiffe 2002). This looks like a good idea at first sight, not only for English and French, the languages the first TAG meta-grammars had been proposed for, but in particular for freer word languages such as German. The German meta-

formalism that connects phrase structure and dependency as we have shown in **Clément et al. 2002**, see Section 4.2.1 .

33 Making the phrase structure comply with generative ideals was also possible in LFG and lead to some debate between the co-founders of LFG, Bresnan and Kaplan (Kaplan and Zaenen 1995).

grammar that I have developed (**Gerdes 2002b**) generates 27,000 different trees for verbs only to capture nominal and verbal valency and the different (topological and categorical) realizations of each argument. Here TAG's system of advantages collapses. Firstly, the actual grammar that does the parsing has lost its explanatory power as the tree fragments are a result of an automatic compilation of the actual simple and manageable rules. The trees themselves are no longer drawn – actually not even looked at – by the human grammarian. Secondly, it is very hard to parse with such large tree sets. The available TAG parsers cannot handle tens of thousands of tree fragments per word form and I have actually never been able to parse a single sentence with my German TAG grammar[34]. We have the contrary case of what happened to Transformational Grammars: TAGs can indeed be parsed in polynomial time but the linear constant is so high that even the simplest sentence results in heavy computing. Finally, this leads to the conclusion that TAGs may actually not be a cognitively intuitive model of Natural Language Processing.

To summarize, TAG has indeed appeared to be a promising formalism, leading the way to reasonable syntactic representations, and, putting practical problems aside, could be a nice model of the syntax module in the brain. The problem is that the formalism simply does not work as expected because it is still stuck in the wrong presupposition that syntax should be formalizable by simple rules. We are making the same error over and over again: We are pushing around the irreducible complexity of language to different parts of the whole model, which gives a short-term pleasure of having discovered a simple rule, beautiful in the linguist's eye, free from empirical intrusion. Then, we proceed to implementing a more complete grammar with the goal of making it actually useful for a real NLP task. And complexity bites us back.

4.2 Generating order

Interestingly, in spite of its name, Generative Grammar attributes structures to given sentences and shows that ungrammatical sentences do not have a derivation. But Generative Grammar never actually generates sentences. Of course, one cannot actually create an infinity of sentences, an infinite set does not have an extensional definition. But one could have started by generating all sentences up to a fixed number of words or up to a fixed depth of embedding, just to get an idea. I do not know of any such work.

There may be multiple reasons for this omission: In spite of linguistics' history as prescriptive grammar, linguistics has actually always nearly exclusively consisted in analysis. Possibly the early prescriptive grammarians did not yet have the formal tools to actually generate a list of (all?) correct sentences. And when the belief in the infinity of language kicked in, it seemed like a futile task. Also, it was not before Chomsky's Minimalist Program that Generativists started to see Language as a correspondence between meaning and text, to use Mel'čuk's terminology (phonetic form and logical form in Chomsky's terms). Chomsky's early grammars are nothing but formal descriptions of sets of grammatical sentences without explicit reference to meaning, meaning only appears in syntactic tests for grammaticality. Syntactic ambiguities were interesting because those sentences have two reasons to be in the set of grammatical sentences, with two different structures. The two

34 But many of the trees have been used in Natural Language Generation projects during my 3-months stay at the DFKI, Saarbrücken, 2000 (Becker et al. 2000).

different meanings were not actually represented beyond the different (deep and surface) phrase structures, the deep structure also projecting down to a logical structure including scope relations. Meaning is also harder to actually put your hands on and it might seem easier to count from one to infinity, corresponding to language analysis than to do the contrary, as in Natural Language Generation (NLG), as Yorick Wilks pointed out. This intuitive image might actually be misleading with its uni-dimensional metaphor of a number beam. It is at least as hard to start analyzing without knowing towards which semantic representation we are heading as to generate text starting from a semantic representation. On the contrary, the result of generation can be tested by native speakers, who may not have much to say about a semantic graph. More generally, "*analysis cannot be done without facing the problem of disambiguation, a problem that cannot be solved (by the speaker or by a formal model) without the use of heuristics based on extra-linguistic knowledge"*[35] (Polguère 1998, p. 4).

So most syntacticians cannot generate, because they do not meet the minimum requirements to start generating a paradigm of sentences: they lack a meaning or other "deep" representation as a starting point, and they lack an operational grammar that can derive a paradigm of paraphrases from the meaning.

To me, writing formal grammars without seeing the different resulting licensed sentences is like mental arithmetic without paper. It is unnecessarily hard. The complex interactions between the different rules are difficult to preview in the mind, and an analysis of a few example sentences seems error prone, because the rule set may be overly restrictive or overly permissive on other sentences.

The next section will show why German word order is particularly interesting, and how I went about actually generating a whole lot of very similar sentences of German; sentences that have the same syntactic representation and "only" differ in their word order.

4.2.1 Ordering German

German word order is quite unique among the world languages because on the one hand, German word order is very free (from purely syntactic constraints), on the other hand, its so-called V2 is enforced with few exceptions (**Gerdes 2005**). V2 means that the verb is strictly in the second position of the sentence, a position called the *left bracket*, following the so-called *Vorfeld* position that can hold any element, a subject, an object, which can even be a subordinated verb, or a modifier. A sentence such as *Yesterday Peter fell* cannot be translated word by word into German. Either *yesterday* or *Peter* can be before the verb, not both words. This structure makes it particularly complicated to describe the German word order in terms of simple precedence rules that were commonly used in formal dependency grammar, but rather requires rules that globally access the word order of the whole sentence. The quite straightforward idea that I developed with Sylvain Kahane is to describe these rules in terms of a rule-based linearization process starting from an unordered dependency tree – a rule system, which we call a *topological grammar* (**Gerdes &**

35 Original text: "*L'analyse, elle, ne peut se faire sans que l'on soit confronté au problème de la désambiguïsation, problème qui ne peut être résolu (par le locuteur ou par une modélisation formelle) sans le recours à des heuristiques basées sur des connaissances extra-linguistiques.*"

Kahane 2001).

It is here that a niche of a short-lived simplification pleasure opened up to me. The aforementioned failure of Generative Grammar to include languages that are structurally different from English (Section 3.2) made it easy to find ways of describing German more elegantly. Inversely, the evident limitations of the dominant phrase-structure-based Generative Grammar scared dependency grammarians away from anything that resembled phrase structure from afar, although prosodic constituents were naturally acknowledged for example in Mel'čuk (1988). Looking at the problem from the perspective of Natural Language Generation pushed us to first go for a very permissive grammar, a shotgun approach: The grammar should include every word order that had been observed, with the goal to add (discursive or selectional) order restrictions at a later stage.

For me, the first preliminary implementation actually preceded a complete formalization: When trying to develop such a formal topological grammar of German, I failed many times to actually make a complete list on a paper of all possible word orders for a given dependency tree. My intellectual limitations obliged me to develop software to calculate all the different ways of saying "Nobody promised this man to read this novel" with the same words in German, which does not sound overly complicated at first sight. What was needed was a kind of manual *bootstrapping* process between the formalization and the software development. The final description gives 99 different word orders out of the 720 theoretically possible (with 6 elements in the sentence, grouping the NPs and "to read" as one unit and 6! =720). Some of the generated sentences are clearly more common than others but for all of them I could find contexts when they appear to be reasonable and very often I could find parallel attested structures albeit with different vocabulary.

So introducing phrase structure as a linearization tool for dependency was novel[36] and surprisingly efficient in producing a specific set of German word orders. But what did that set actually represent? The satisfying part of the grammar we described is that it describes the backbone, the hard part of the German word order, we will get to the soft tissue in the next section. The V2 order and the verb cluster are reasonably simple to describe and this simple beauty leads me to keep up my faith in the actual (neuronal) existence of these rules. But even if the most efficient neural representation of German has nothing that resembles topology, these rules have a pedagogical aspect to them as every learner of German has to get to know them at some point. So maybe they are just good answers in the very specific framework of a scientifically trained linguist or language learner, as we will discuss in the concluding Chapter 7 .

Nevertheless, topological linearization grammars have shown to provide similarly interesting word-

36 During my stay at the DFKI in Saarbrücken in 2000, I met Ralph Debusmann, who was also working on a formalization of the classical notion of topological fields for German. I had already tried to add those fields as nodes to the TAG trees in **Gerdes 1998** and Ralph was working in the framework of constraint grammars. This resulted in two topological field formalization papers at the ACL 2001 conference: Duchier & Debusmann 2001 and **Gerdes & Kahane 2001**. The idea was somehow due and up for grabs: Formal grammars were increasingly looking into different languages including German, the topological model of German was well-known and widely used in the teaching of German, and Kathol (1995), based on work by Höhle (1986), had already used the field names in his Phd on German HPSG (Head-Driven Phrase Structure Grammar) – Kathol's book had just been published as a book in the same year (Kathol 2000). This is a tiny but nonetheless interesting example of the cumulative structure of scientific progress, where the individual contribution seems pre-determined: The scientist just perpetrates the "discourtesy of having usurped it first" as Borges (1923) put it. See Section 6.6 for the complete quote and further discussions on the notion of *novelty*.

order descriptions for a whole range of typologically diverse languages such as Arabic, Chinese, French, Greek, and Korean. The topological model of the Chinese sentence (**Magistry & Gerdes 2009**) is particularly interesting, not only because it contradicts Humboldt (1836) who said that "the Chinese method is too feeble in conveying the feeling of sentential unity into the language", but mainly because with few rules, we managed to predict complex structures such as serial verbs and the "ba" and "bei" constructions, the latter being sometimes referred to as "passive". Our rules then correctly predicted "for free" complex constraints on left dislocation, very common in spoken Chinese but much less frequent in written texts. For French, we discussed in **Gerdes et al. (2005)** the "descriptive adequacy of topology", because a satisfying topological grammar does not produce intuitive constituents. Topological grammars have two sides: They are applied grammars for use in generation systems and at the same time they are a way of understanding word order constraints. Yet, they remained ambivalent concerning whether the topological structures themselves, that are constructed in the linearization process should be given a theoretical status – whether they actually "exist" or are mere artifacts of the process of ordering words.

A topological linearization grammar has in common with a Generative Grammar that it studies potential, not actual word orders. Yet, it is very different from Generative Grammar: Firstly, it is implemented, at least from syntax to topology[37], and the generated sentences can actually be looked at. Any speaker of the language can look at a generated sentence and can attempt (and possibly fail) to come up with a context that licenses the sentence. Compare that situation with Generative Grammar, where all we get is a complex phrase structure tree, usually higher than wide, on top of each sentence; a tree whose value is purely theory internal. And secondly, the topological linearization grammar is part of a sequential (or language) model, also called pipeline model, linking meaning with sound output, i.e. a model ranging from semantics to syntax to topology to phonology to phonetics, where each meaning goes through the different transformations to become a sound output. In contrast, originally, Generative Grammar only tried to describe the set of grammatical sentences, while modern versions add more and more language data to the phrase structure, ranging from the phonological output to semantic scopes.

This pipeline model itself is another example of a pleasant over-simplification. When trying to fine-tune the rules, one stumbles upon relations that seem to take shortcuts. One example is how prosodic groups depend on the sentence's theme-rheme structure (**Gerdes & Kahane 2000**), which in the Meaning-Text Theory is already present in semantics. This requires to carry the theme-rheme structure all along the pipeline. This theme-rheme structure can also interfere in the word order itself, which will be discussed in the next section. There are even a few examples of two elements occupying the place before the verb together, apparently violating the German V2 rule. This can be explained away by means of the theme-rheme opposition intervening in the semantics-syntax as well as in the syntax-topology interface, as I have shown in my paper presenting *Tree Unification Grammars* (**Gerdes 2004**).

Having described the hard topological constraints of German, the next natural step is to dive into

37 For the reverse direction, from topology to syntax, we have developed an algorithm that can be constrained to be polynomial in **Gerdes & Kahane (2006)**, Moreover, not only for the syntax-topology interface but even for Mel'čuk's whole Meaning-Text Model there are a few existing implementations, rule-based and statistical, in particular in Leo Wanner's group (Wanner et al. 2010, Bohnet & Wanner 2010).

the generated sets of sentences and to formalize the preferences of one word order compared to another. That's where things are toughening up.

4.2.2 Wobbly word order

One of my first jobs in linguistics was an internship at a company, LexiQuest, where I had to improve a German weather forecast generation grammar developed by Leo Wanner to be integrated into the text generator developed by José Coch (1996) and integrated into the multilingual MultiMétéo system (Coch 1998).

My job consisted in improving the German grammar that had been judged by native users as producing unnatural sentences. The following example is a perfect real-world example (taken from wetter.com):

(1) In Paris strahlt am Morgen die Sonne und die Temperatur liegt bei 12°C.
in Paris beams in_the morning the sun and the temperature lies around 12°C

'In Paris there will be sunshine in the morning and temperatures will be around 12°C.'

Just looking at the first part of the sentence "In Paris strahlt am Morgen die Sonne" 'In Paris the sun shines in the morning' we have 3 verbal dependents: "in Paris", "am Morgen", and "die Sonne", giving *3!=6* grammatical orders of these dependents:

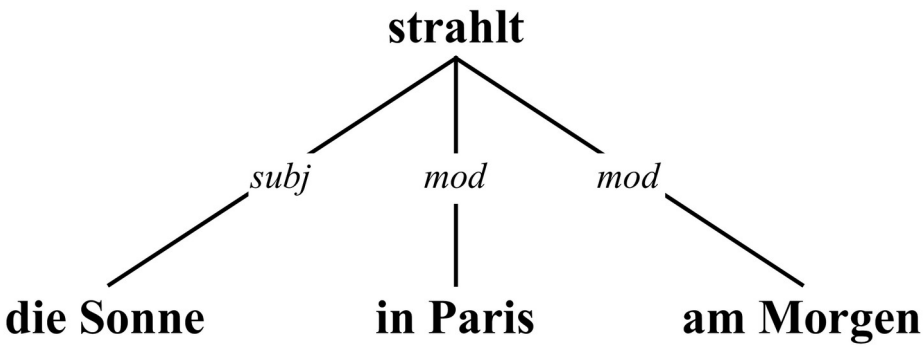

Illustration 4: Dependency tree for "In Paris strahlt am Morgen die Sonne"
'In Paris there will be sunshine in the morning'

a) In Paris strahlt am Morgen die Sonne. 'In Paris there will be sunshine in the morning', 'The morning will be clear in Paris.' (the intended reading)

b) ?In Paris strahlt die Sonne am Morgen. (acceptable with focal accent on "Sonne" 'sun')

c) Am Morgen strahlt in Paris die Sonne. (not as nice as *(1)a*) but good if contrasted, for example followed by 'and in the afternoon...')

d) ??Am Morgen strahlt die Sonne in Paris. (acceptable with focal accent on "Sonne" 'sun')

e) ????Die Sonne strahlt in Paris am Morgen. (unacceptable except in some language teaching or correcting context)

f) ???Die Sonne strahlt am Morgen in Paris. (unacceptable except in some language teaching or correcting context)

Before my intervention, MultiMétéo's original generation grammar chose one of the orders by a rigid "subject object modifier" order, (1)e, the worst of the possible orders. (1)e is a semantically bizarre sentence whose first interpretation is something along the lines of 'In Paris, sunshine happens in the morning', because the best interpretation takes the last element, 'in the morning' to be the rheme of the sentence, the preceding part being the theme: 'sunshine in Paris'. Although this seems to be only a different theme/rheme distinction of the same predicate structure, the meaning appears to be completely different. Although 'in the morning' can mean 'every morning' or 'this morning', under the title of "tomorrow's weather forecast for Paris" this ambiguity could be expected to be easily resolved. Nevertheless, even in this very specific and narrow context, sentence (1)e has a very strong tendency to receive the unintended interpretation, that actually has different truth conditions stating that in general, Parisian sunshine is possible (only) in the morning. And although, of course, German readers can recover the actual meaning, the sentence in this form is highly disturbing, as the sentence reveals its non-human origin, making the output considerably less usable. We could speak of a lack of "naturalness" to use Sinclair's (1984) term that he opposed to a more absolute grammar-based "well-formedness" but order (1)e is not just unnatural: It induces the reader to a wrong interpretation.

To remedy these problems, I wrote a whole bunch of ad-hoc ordering rules for the German MultiMétéo generation grammar, making the result much more human-like. But the effort left an unsatisfactory aftertaste as most rules seemed highly idiosyncratic and no systematic evaluation was possible, lacking actual German forecast corpora as well as the possibility to systematically present to German natives the resulting data in its context.

So many factors seem to interfere: the enunciation context of the sentence (on a weather forecast site), the discourse context (the preceding sentences), the weight of the words and phrases, pronominalization, and the idiomaticity of the dependency relations.

To illustrate the latter point, firstly it is important to note that "strahlen" 'to beam' is a collocate, more precisely the light verb with the lexical function $Func_0$, of "Sonne" 'sun' (Mel'čuk 1998). This causes the predicate to be "Sonne strahlen" and putting "Sonne" in the end gives the desired thetic (i.e. all rhematic) interpretation of the sentence as the predicate reaches from the left bracket (i.e. the V2 position, see Section 4.2.1) to the end of the sentence. Yet, it is equally important to note that in German, the sun "beams" in the morning, i.e. "Sonne strahlen" actually nearly exclusively happens in the morning. Google gives me 9240 results for "strahlt am Morgen" 'beams in the morning' and only 70 "strahlt am Abend" 'beams in the evening', although overall, "am Abend" is more frequent than "am Morgen". The idea is that the light verb "strahlen" 'to beam' is something the sun does in the morning.

It is tempting to look and fall for some kind of "logical" explanation: In the evening the sky is often hazy and the sun cannot 'beam' anymore; the German sun prefers to simply shine "scheinen" in the evening. Note that the whole word set of order constraints is different when using the alternative light verb "scheinen". So the complex predicate that we have in this example might include the

whole “die Sonne strahlt am Morgen”, which could be the reason for the high acceptability of (1)c. And this might also explain the still quite high acceptability of (1)b: “am Morgen” can move in the “tail” part of the sentence (the thematic part that is not accented, Vallduvi 1995) because “am Morgen” is actually semantically expected as part of the predicate. But this interpretation cannot be sufficient because of (1)e: If the goal for a thetic sentence is to make the predicate reach to the end of the sentence and “am Morgen” is part of the predicate, then this word order should be more or less acceptable, too. It is however the worst of all possible word orders.

Note that I am not defending a radical constuctionist syntax that would rid itself of global categories and syntactic functions altogether (Croft 2001). Of course, all words in (1) behave normally within what is expected from their category and syntactic function. I would not even like to call those order constraints a *construction* in the sense of Fillmore et al. (1988). There is nothing much going on between the meaning and the form in this example: The collocation “Sonne strahlen” put aside, everything else seems compositional, word order constraints apart. I would not even be certain which lexemes to include into the description of such a hypothetical construction: Is “am Morgen” part of it or not? It seems unlikely that there is a second configuration that has degrees of idiomaticity and entropy so similar to the present case that the linearization behaves in the same way. We seem to have come across the oxymoron of a *compositional set phrase* here, that, if used in the given context of a weather forecast, does not license any order variation.

One is inclined to simply conclude by paraphrasing Ibn Mada : Word order (1)a is how the Germans say that the morning will be clear in Paris.

Still having the beauty of simplicity in my mind, I took away from this experience only one lesson about German: The verbal dependent order is something to avoid. Henceforth, I decided to restrain my research to the quest of constructing the backbone of topological grammar – the rules that describe sentences for which I can find at least one interpretation and, even better, for which I can even find a parallel attested real-world example. And I do not try to decide which of these orders is best in a given context. Stefan Müller (1999, p. 171-172) resigns similarly by stating that his HPSG grammar can only be used for analysis because if used in generation, it would generate many “unaccepatable and marginal sentences.” – the corresponding generation grammar that would be able to order German verb dependents is simply being put off for ever and a day.

There is nonetheless a second lesson to be learned, a lesson in linguistic methodology: Generating paradigms of sentences gives you the necessary overview over any grammar rules one might propose. Without this tool, one not only misses the right rules, one easily misses the fact that there might be no simple rule at all.

5 On corpora

5.1 Defend us Lord from complexity

While the non-generating Generative Grammar dominated linguistics, corpus linguistics did not disappear completely. Corpora survived in the study of linguistic variation (Léon 2007), for example in France with Blanche-Benveniste's Aix School studying spoken language (Blanche-Benveniste et al. 1979, 1990) and in Britain around the creation of the British National Corpus (Leech 1993), in language acquisition (MacWhinney & Snow 1985), and of course in more remote fields such as phonology and discourse analysis. This survival of corpus studies in the sociolinguistic and acquisition niches can be attributed to the evident uselessness of asking infants or minority groups how their grammar functions. However, it should not go unobserved that following Chomsky, even the standard speaker does not have direct access to their grammar as they may not be aware of taboos and other non-linguistic constraints that govern their language. Chomsky defends what Devitt calls the Representational Thesis: Humans have an internal representation of (generative) grammar rules but "a person could be competent in a language without representing it or knowing anything about it: she could be totally ignorant of it" (Devitt 2006). So in a sense, the introspection leading to the discovery of the grammar rules is reserved to the elite of well-trained, not to say brainwashed, linguists.

The survival of corpus studies in fields such as sociolinguistics and acquisition was also conditioned by the fact that these researchers refrained from claiming anything about grammar, reminiscent of the possible coexistence of science with a religious worldview as long as science has no claims on morals (Harris 2011).

The coexistence of Generative Grammar and empirical linguistics lasted surprisingly long, fostered by the aforementioned epistemological reasons limiting corpus linguistics, ranging from the technical capacities of the involved scientists to the desire to find simple formulas just like in physics. The opposition between simple and beautiful models on the one hand and methods of handling empirical data on the other can even be found in mathematics: It is the opposition between algebra and (numerical) analysis. Personally, I was attracted to the algebra side of mathematics and to me, when moving over to linguistics, the questions asked by generativists seemed naturally right. It is also easy inside formal linguistics to not even hear about the empirical approaches that were developing in parallel because it seemed that they asked completely unrelated questions.

Linguistics took a long time to acknowledge that nothing will be the same anymore with the emergence of a resource of textual data as enormous as the Web. In the 1990ies, the new possibilities of large corpora were explored first in computer science, for example in the field of speech recognition, but also in sociological discourse analysis. It is from these very different fields that corpora came back to formal syntax, where they had only survived in some niches.

It is also too tempting to qualify the astonishingly large communities of textual statistics (the JADT community in France or the Corpus Linguistics community in Britain) as heterogeneous crowds consisting of a few programming gurus developing the algorithms and programs, and users who do

not understand the results they are seeing when they put their texts into the tools because they have no idea of statistics. Yet, in a sense, the textual statistics communities kept their scientific instincts intact: There is new data, there are new tools, let's try out what we can do – a reflex that had perished among syntacticians. The Big Data revolution only hit linguistics when the tools were getting more powerful and sophisticated, and they were coming to the well-preserved kernel of linguistics: syntax.

5.2 The wow-effect of statistics

I remained clean of statistics until my appointment to the Sorbonne Nouvelle in 2006 where my colleagues André Salem and Serge Fleury developed tools of textual statistics. I still remember my awe when a simple bitext, a text with its translations aligned on the paragraph level, allowed getting translations of most words. No annotation, no syntax, no analysis. Just raw aligned data.

Translations are a remarkable thing to begin with. I am not the only linguist who has been drawn into linguistics by the fascination of translation. The dream of machine translation is as old as the dream of the computer. It is the application par excellence of a smart machine.

A translation is just another type of paraphrase (Mel'čuk 1988b). The same meaning can be expressed by different means, in a different code. It is worth marveling at the sheer fact that every text can be translated. The text may get longer, some puns may need explanations, the vocabulary will get poorer – we are looking at what is sometimes named "Translationese", a recognizable language style (Baroni & Bernardini, 2005). But the meaning can always be rendered in the other language, a reminder of the early Wittgenstein's equation of the speakable and the thinkable.

This idea that somewhere in the words of the translation the original meaning has to be encoded explains why statistics can make such good predictions of translations of words, in particular if the corpus is large and homogeneous. So all we need is a large list of couples of paragraphs, each couple consisting of a source paragraph and a target paragraph. Couples of sentences will do even better, but the more source and target languages are typologically different, the harder it is to make a sentence alignment, i.e. to find linearly ordered couples of source and target sentences because sentences will be translated in a different order (or a translation partner may be entirely absent). Paragraphs, however, are more universal in that they describe borders of chunks of meaning. Even for paragraphs, the writing conventions will make the translator split the text differently, but only very rarely will the translation invert paragraphs.

Such aligned bitexts are a valuable resource today as they provide the so-called "translation model", the backbone of all statistical machine translation systems (SMT). The most complete online bilingual concordancer is linguee.com, and this database allowed the company to build the currently best machine translation algorithm for general texts (see for example Isabelle & Kuhn 2018 for a recent comparison of French to English Machine Translation). I have worked on the construction of bilingual resources for quite some time. For instance, together with Kyungmin Shim, a PhD student whom I codirect with Bernard Bosredon, we have built a French-Korean bilingual concordancer, the only existing one to my knowledge, available at https://kofr.ilgpa.fr containing around one million words on each side (Shim 2017)

The choice of the correct translation for a single source word *S* is just a sub-task of the SMT translation algorithm. The setup is quite simple: We take a large list of bilingual couples of paragraphs and select the paragraphs that contain *S*. Then we select the set $TP = \{tp_1, tp_2, \ldots tp_n\}$ of all corresponding paragraphs of the target languages, i.e. all paragraphs that are aligned with the selected paragraphs of the source language that contain *S*. The selection of paragraphs of the target language *TP* is an arbitrary collection of text – with one important exception that distinguishes this selected text to the other unselected paragraphs: The selected paragraphs *TP* all contain translations of *S*. Now all we have to do is to measure which words appear astonishingly frequently in *TP* compared to the whole target text (Zimina 2004). This is not the same thing as measuring frequency itself. The most frequent word will most likely be *the* in just about any selection of texts in English. To assess the degree of astonishment of over-representation of a word in a given text, different measures have been proposed, ranging from simple frequency to maximum entropy based measures. The notion of *specificity*, a measure introduced by Lebart & Salem (1994), is the logarithm of the p-value of the Exact Fisher Test (Fisher 1922), based on the cumulative hypergeometric distribution. The higher the absolute value of the specificity of a word, the less likely it becomes that the observed frequency is due to pure chance – and the more likely it becomes that some force other than pure chance was behind the observed frequency. Since the only discriminating point of *TP* compared to the rest of the text was that it contains translations of *S*, the word that is the most astonishingly over-represented is probably *T*, the translation of *S* (or a part of the translation of *T*, if there is no one-word translation for *S*). The translation is simply the word of *TP* with the highest specificity.

This does not always work perfectly and sometimes *S* has, for example, two equivalent translations T_1 and T_2 with the same frequent modifier *M* and the hit parade of specificity starts with *M*, T_1, T_2. But if the corpus is sufficiently large, there should be cases of T_1 and of T_2 without M, and the results become remarkably good. Yet, the statistical measures themselves remain difficult to access for two reasons.

Firstly, every statistical test needs a null hypothesis, the default case where no relation exists. And the null hypothesis of word distribution underlying specificity measures are clearly counter-intuitive. Supposing a hypergeometric distribution of words is tantamount to considering that writers have a fixed bag of words, and writing consists in distributing these words. The null hypothesis says that without special reason, the words are distributed equally over the pre-defined sequences that are measured: Sentences, paragraphs, or texts. We measure to which extent it is possible for this null hypothesis to be true. This implies that a word that is taken out to be used is less likely to appear again. The contrary has been shown for the lexicon and syntactic structures: The more someone hears a word or a construction, the more likely becomes the re-use of this word or construction (Bod 1998).

Secondly, different statistical measures can and should not be compared based on their null hypothesis. They can only be judged on their usefulness, i.e. to which degree the results correspond to the expected values. And many different measures do give good results. As André Salem explained to me, it doesn't matter how precisely you put a horizontal plane over France, you will always measure that the Mont Blanc is the highest mountain. So when textual statisticians try out

their tool on, say, the works of two authors, they will obtain a list of the most over-represented words in the works of *Author 1* compared to the corpus of *Author 2*, and they will like the measurement when it corresponds to their expectations. And the only way to asses that, for instance, specificity is a better measure than maximum entropy is the higher satisfaction of the researcher. This can be caused by specificity's lower frequency of function words, called "stop words", at the top of the specificity list. In maximum entropy measures these words are usually sorted out by means of a pre-defined stop-word list and completely ignored because otherwise they pollute the results, and specificity does not need this list of stop words. But how do we know that this is actually better? Maybe the important differences between *Author 1* and *Author 2* lie precisely in the different usage of function words (which may also appear, but less prominently, when using the specificity measure)? But such a difference has never been used in the comparison of two authors – not because it is useless to compare function word usage between authors but because the high frequency of these words are undetectable without tools.

It is a bit like declaring that the important difference between say goats and sheep is the shape of the horns, because for some reason that is what we remark most – it is what we are able to see easily. And then we use this gauge for measuring the difference between other animals, too, although our tool has been fine-tuned for horn shapes. For someone like me coming from the simple beauty of Algebraic Geometry, there is something deeply unsatisfactory about this way of choosing a work method. But, as I shall argue in the conclusion, precisely this quest for intellectual satisfaction is what science will be all about when it will no longer be needed for engineering.

5.3 My corpus tools

These considerations kept me from tampering with the actual statistical measures and brought me to developing tools simply based on a standard implementation of specificity as one of the measures that give satisfying results. The tools I developed focused rather on the corpus creation from two sides: on the alignment process and on the corpus collection from the Web.

At first, I built a simple online tool for manually aligning source and target paragraphs of a bitext (https://elizia.net/alignator/). There is a whole community around statistical tools that mostly try to take educated guesses on how to sentence- or word-align a bitext, commonly either with the assumption that there are some cognates, i.e. similar looking words, between the languages or with some small dictionary to start the process. I was interested in a tool that does not need any prior dictionaries or similarities between the languages, as in the case of aligning French and Chinese, where absolutely no similarity can be presupposed (e.g. even numbers and punctuation are different Unicode code points). So I showed in **Gerdes (2008a and 2008b)** that it is sufficient to know that two texts are the translation of one another to make the hypothesis that words or morphemes will exist that appear at similar places in both texts. The distance from one appearance of the word to the next (in percentage of the whole text) gives some kind of fingerprint of a word. It is then possible to predict that some couples (s_i, t_i) are very likely to be mutual (partial) translations which can be computed by means of a Dynamic Time Warping algorithm. Once the most likely couples have been established, these couples can be used as preliminary anchors for a first coarse alignment, which allows in turn to start a bootstrapping process based on the specificity values described

above. This works astonishingly well for French-Chinese text alignment, although the first anchors are mostly not complete Chinese words but rather single characters, for example French *chaise* 'chair' is aligned with 椅 *yǐ*, the first character of the Chinese word 椅子 *yǐzi* 'chair'. The results can be improved by grouping French words by a simple variant of the Levenshtein distance (the Jaro-Winkler distance, where word beginnings are less flexible than word endings). The problem with all that, however, is that the determination of the concrete thresholds of what is a good word couple and which kind of pre-treatment is best, depends not only on the language couple but also on the text genre or the translation quality, for example the frequency of translator's notes. Although the basic algorithm is stunningly powerful, it takes lots of fiddling with the actual texts to get optimal results, and, most importantly, it is hard to take home generalities about bitexts.

The second corpus tool that I developed is the *Gromoteur* (**Gerdes 2014**), a Web spider specialized in the linguist's needs and with a user interface that cuts up the configuration process into steps of an interactive wizard. Contrary to other Web spiders, it allows a user to collect and clean whole websites without having to access the command line. Although the tool was originally conceived for my own needs as well as pedagogical use in my NLP introduction classes, the overwhelming feedback that I received from the linguistic community, with many special needs and surprising bug reports, made me spend much more time on this tool than scientifically reasonable. It was also a good lesson in the difficulties of developing software with a wide distribution – and most importantly keeping it up to date with constant changes in the operating systems. Contrary to most other tools that can be put on a server and accessed by means of a simple browser, it is impossible to conceive an actual Web spider as a web service because a single server that downloads whole web sites for many users will quickly be blocked by those servers (because the repeated spidering will be considered as a DOS attack). The project is still alive, and together with Serge Fleury and Gaël Guibon, I am currently trying to develop a Javascript based tool that will run inside the user's browser, thus eliminating the problem of different operating systems and allowing for the integration of different statistical analysis tools. Computing the specificity values of a whole corpus is computationally expensive and was unimaginable to realize in Javascript only a few years ago, even Python was much too slow, and the specificity computation is commonly programmed in C++. It has become possible thanks to the recent important browser speed improvements in the interpretation of Javascript.

5.4 In defense of stamp collection

The structuralists' insistence on the infinity of Language can only be explained epistemologically: As pointed out in Section 3.3.1 , the very high number of possible combinations seems infinite to a human mind bereft of technical memory storage and search tools – we shall return to different interpretations of *infinity* in Chapter 6 . Moreover, it was necessary to discredit corpus use as a linguistic method, and collecting language data seems less expedient in face of infinity.

Yet, it is important to point out that the finiteness of language does not mean that a sheer list is the best language model. The human brain does not use lists – it can understand sentences it has never heard. But it may have more exceptions and frequency information than general rules. Lists are not only cognitively implausible: For any other imaginable application of linguistic discovery, such as

scientific satisfaction, Natural Language Processing, first and second language teaching etc, a "zipped" version of the corpus, i.e. the information theoretically most compressed form of the data, prevails as the goal of linguistics. The question is only how to obtain this minimal set of rules.

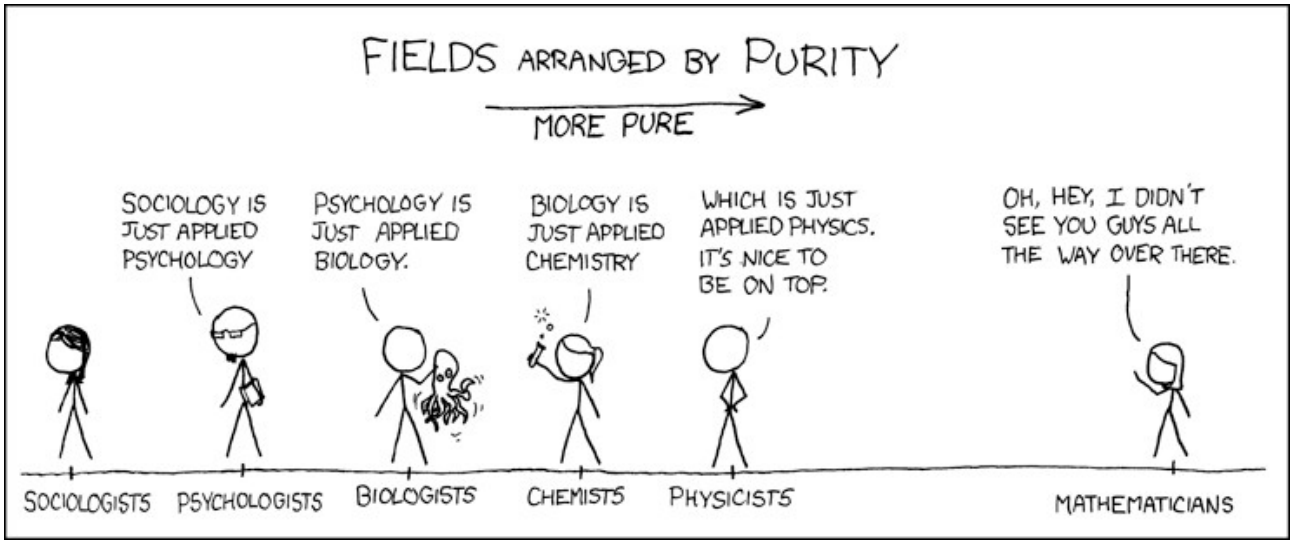

Illustration 5: https://www.xkcd.com/435/

"All science is either physics or stamp collecting" is what Ernest Rutherford declared in order to put his science on a pedestal way above mundane classification tasks. But is it actually possible to classify without at least a glimpse of a general rule?

Is there a fundamental difference between stamp collection and treebank[38] annotation? Annotating a sentence is a bit like describing a butterfly: blue wings, black eyes, feathery antennae, etc. We will show in Section 6.3.1 that transcription can be seen as a categorization into a very large but finite number of categories, the words of a language. If we move on to treebank construction, the task can essentially be seen as giving examples of what we want to talk about: subjects, objects, appositions, … And Hockett (1948: 269) already asserted that "*the task of the structural linguist, as a scientist, is [...] essentially one of classification.*" And he goes on: "*The purpose, however, is not simply to account for all the utterances which comprise his corpus at a given time; a simple alphabetical list would do that. Rather, the analysis of the linguistic SCIENTIST is to be of such a nature that the linguist can account also for utterances which are NOT in his corpus at a given time. That is, as a result of his examination he must be able to predict what OTHER utterances the speakers of the language might produce, and, ideally, the circumstances under which those other utterances might be produced. […] In theory, at least, with a large enough corpus there would no longer be any discernible discrepancy between utterances the linguist predicted and those sooner or later observed. [… The linguist's criterion is] accuracy of prediction beyond one's corpus, and communicability of the basis for such prediction to others.*"

Hockett not only explicitly linked classification and prediction. It is fascinating that he already saw the accuracy of the system on a par with its communicability. Being able to communicate, to

38 A *treebank* is simply a syntactically annotated corpus. It has received that name because syntactic structures are conjectured to be simple trees (rather than more complex directed graphs). We have grappled with that restriction to trees and proposed extensions in different papers (**Gerdes & Kahane 2009, 2011, 2015, 2016**).

transmit, to apply the discovered rules is absolutely central to linguistics. And in today's world of machine learning, we no longer have to oppose descriptive and explanatory science. Machine learning is about leaving it to the system to find the best compression of our data, the best predictive accuracy. We collect the stamps, the machine finds the underlying system. It picks up the rules behind our description, behind the annotation. This allows us to group the data, to automatically find many more similar examples for any configuration, and finally to take distribution measures on our phenomena. For any construction that we can identify, the rules thus give us the complete paradigm of similar constructions found in the corpus. But how to get the rules that please us, the rules we want to communicate? The answer to this question is central to today's linguistics and must be empirical, i.e. it must rely on access to large amount of textual data by means of tools. Before attempting to give answers to this question in Section 5.8 , we will first list the whole range of ways of making use of massive amounts of language data, of doing corpus-based linguistics:

5.4.1 Discovery of regularities by annotating data

Any linguistic annotation process allows confronting our ideas to actual data. This triggers different bootstrapping loops: The annotation scheme (segmentation, categories, functions) is revised if it cannot be applied systematically and coherently – applied by the same annotators, by other annotators that follow the annotation guide, and finally by machine learning systems trained on the first annotated data. Practicing annotation can profoundly change our view of what is important in the data – the frequency of the phenomena to annotate dictates what we have to look into. We have experienced such a change of interest in the various annotation projects I was involved in: Working on Spoken French in the two French National Research projects (ANR projects) Rhapsodie (**Lacheret et al. 2014**) and Orféo (Benzitoun et al. 2016, annotation guide **Kahane et al. 2018**), we were forced to look into paradigmatic phenomena such as disfluencies, elaborations, co-constructions, and coordination, leading us to the notion of *piles*, see Section 5.6 , **Gerdes & Kahane 2009**). The Procore project (a two-year Franco-Hongkongese Hubert Curien Partnership) that I led from 2016 to 2017 together with John Lee from the City University of Hong Kong on the Semi-automatic Creation of a Parallel Treebank of Cantonese and Mandarin (**Leung et al. 2016**), has made us more aware of the fact that many Mandarin prepositions, which had originally evolved from verbs but are now fixed and no longer function as verbs, have Cantonese equivalents which are still showing verbal behavior, such as accepting an aspect marker. This leads to important differences in the distribution of parts of speech between the two neighboring languages (**Wong et al. 2017**). The ongoing annotation work on Naija in the NaijaSynCor ANR project (**Courtin et al. 2018**) has confronted us with the difficulty of distinguishing categories and determining the root node, the central syntactic element of a sentence, independently of communicative considerations, in this evolving isolating language without a grammatical tradition.

5.4.2 Testing hypotheses on corpora

We can verify the existence or absence of linguistic phenomena. This can be done by simple concordancers and frequency counts on raw or annotated data. Although this constitutes the most basic approach to textual data, it reaches its limits in the difficulties of actually using the tools, in particular the often complex search tools for annotated data, and of interpreting the observed

numbers (See Section 3.4). This approach was central to our work on topological fields and the determination of the scope of word order variation that we want to include in our German grammar (**Gerdes & Kahane 2001**). Since annotated corpora of German and other languages that we worked on were still in their infancy, we principally relied on manual variations of queries to Web search engines.

5.4.3 Study of correlations on double annotated corpora

A slight step up is the corpus study of correlations between two independent aspects of language. This requires corpora with a double annotation, such as syntax and prosody, which was the goal of the Rhapsodie, Orféo, and NaijaSynCor projects. Our work on the relation between prosodic and syntactic constituents (**Gendrot et al. 2009, 2010,** and **2011**) examined for example which kind of syntactic structures of French lead to prosodic breaks, marked by vowel lengthening. We showed that at least for normalized journalistic pronunciation, cognitive or articulatory constraints such as the preferred grouping of 7 syllables actually predict best where the breaks occur, see Section 3.3 .

5.4.4 Discovery of regularities by textual statistics

It is possible to start with an elementary hypothesis such as “there is a difference” and get help from statistical measures to actually put a name on this suspected difference. This approach is frequently applied in textual statistics (Lebart & Salem 1994). For example, my web spider and textual analysis tool *Gromoteur* (**Gerdes 2014**) that I have presented above (Section 5.3) allows to study the differences between two news websites, an exercise that we frequently proposed in the Digital Humanities and Web Corpora classes I taught with Serge Fleury and Isabelle Tellier: Students choose two news websites where they suspect thematic differences, for example based on the political orientation of the newspaper’s owners. Then we collect around a thousand articles from each site. The statistical analysis allows to list the specific words (see Section 5.2) of each site that then need to be interpreted either as artifacts of the collection process or as indication of a particular interest and commitment of the newspaper. In the same vein, I worked with Miao Jun on the computation and visualization of the translator’s style (**Miao & Gerdes 2011**): If a book has been translated independently by two translators, we can suspect that the different word usage in the two translations cannot be due to the content, which should be identical, but to stylistic choices that the translator has made in the translation process, revealing his or her particular interest in the original text.

5.4.5 Typometrics

The emergence of treebanks following the same annotation scheme for a wide range of languages in the Universal Dependencies project (UD, Nivre et al. 2016, **Universal Dependencies Version 1.2** to **Version 2.2**) allows new ways of verifying existing claims on languages’ syntactic proximity and family relations (**Chen & Gerdes 2017** and **2018**). But it also allows discovering yet unknown correlations between languages. In our recent work on what we call *typometrics* (**Gerdes et al. submitted**) we use a specific type of two-dimensional scatter plot that permits studying word-order behavior, for example the percentage of right-branching subject in relation to the percentage of

right-branching objects, for all 70 languages in the UD family. We can even go one step further and compare correlations. For example, we show that in human language, the distribution of the directions of all dependencies compared to adpositional objects (i.e. the mutual position of the preposition and its argument, e.g. *for→Peter* where the link goes from left to right) is very similar to the distribution of the directions of all dependencies compared to auxiliary-main verb relations (*having→yawned*). An absolute typological universal can now be seen as a special case of a typometrical distribution.

5.4.6 Assessing annotation regularity through statistical parser performance

Another interesting way of deriving syntactic knowledge from treebanks consists in the usage of statistical parsers. The basic idea is that coherently annotated data is easier to learn than incoherent data. Equally, local rules are easier to learn than long-distance relations. It is thus possible to compare different annotation schemes by comparing parser performance on the same texts. Of course, we always also measure the limitations and tendencies of the parser itself, and moreover, neural network based parsers are increasingly capable of discovering complex relations that are distributed widely in a sentence. Yet, if different parsers give better results on one annotation scheme over another, we can conclude that this annotation scheme is preferable in some sense, even more so if the argument can be corroborated by theoretical considerations about the annotation scheme. This is what we have done in order to develop and promote the SUD annotation scheme variant of UD, see Section 5.7 and **Gerdes et al. (2018**).

5.4.7 Measuring syntactic distance through parser cross-training

Parser performance can also help us to assess syntactic proximity between texts or even between different languages. The idea is that if a parser trained on text A gives better results on text B than on text C, text A and B are syntactically more similar. A systematic cross-training and testing of a set of texts gives a distance metric between the texts, or, strictly speaking, between the treebanks that can be extended to the texts themselves under the condition of a homogeneous annotation (**Guibon et al. 2014 and 2015**). See Section 5.7 for further details.

Note that these preceding two methods of parser use give a new twist to the increasingly central role of shared tasks and their evaluation process in NLP research, “often suggesting that the reference solution is an objective reality”[39] (Poibeau 2011). In these methods, on the contrary, the evaluation techniques are used only indirectly, in order to compare texts and annotations, not with the goal to improve the parsing techniques themselves.

We will now dive deeper in what can be learned from syntactic annotation.

5.5 Syntax Ltd.

When doing syntax, in particular corpus-based syntax, we have to define the lower and the upper limits of our annotation. This means that we have to define the building blocks of syntactic structures as well as the maximal construction we want to achieve – we have to determine when to

39 My translation of “*laissant souvent penser que la solution de référence constitue une réalité objective*” (p. 34)

stop building. Both limits are hard to define even on written texts but the definitions are clearly less trivial on spoken data.

5.5.1 Upper limits

The upper limits of syntax are particularly difficult for spoken language, understood here as a continuous transcription of words without delimitations. Put simply, in spite of doubts about some punctuation signs (semicolons, colons, quotes indicating direct discourse, …), a syntactic structure on written texts should combine words between two periods. For spoken data, it is necessary to define the position of sentence-delimiting signs on transcriptions in a way that the sentence segmentation is reproducible. It turns out that this is a very hard task: Humans frequently disagree and machines are facing a hen-and-egg problem: They need to analyze the sentence to know where it stops but most current parsers need a finite list of words to start analyzing. This is surprising because it seems that we have the whole Web of one language that we could train a statistical sentence segmenter on. The conundrum is that this task is just as difficult as semantic parsing without any syntactic structure: The (neural) semantic parser and also the sentence segmenter have to somehow come up with an approximation of the syntactic rules themselves. The obvious step out of this loop is to develop incremental parsers that at each step give the probability to have reached the end of a sentence. There is work in this direction but it seems quite disconnected from the common parsing community that organizes shared tasks on written data with delimited sentences (see for example Honnibal & Johnson 2014).

Epistemologically, it is interesting to observe the gut reaction of descriptive linguists working on real-world data against the notion of “sentence” (see for example Berrendonner 1990 and Pietrandrea et al. 2014). Quite a few spoken corpora do not put any punctuation at all, such as the Childes corpus (MacWhinney & Snow 1985), the corpora from the Aix School, or the TCOF corpus (Benzitoun et al. 2012) and settle with speech turns. Others put specific signs such as “//” to delimit the upper limits of syntax – a solution that we have adopted for Rhapsodie (**Benzitoun et al. 2010**), for Orféo, the follow-up project of Rhapsodie with much more data and a written data component for comparative analysis (Benzitoun et al. 2016), as well as for NaijaSynCor, the ongoing treebank development project of the Nigerian Creole (**Courtin et al. 2018**). Avoiding the dot symbol and the word “sentence”, we settled for the notion of *illocutionary unit*, which has the clear advantage to point to the central defining notion: Illocution. We call *illocutionary unit* a sequence of words that can stand alone as an illocution. We could say with Humboldt (1836, p. 128) that such a unit “constitutes a closed thought in the speakers’ mind”, i.e. it can be pronounced individually with the same meaning as in the context we found it. Of course, this definition directly depends on our understanding of “meaning” and in particular “same meaning”. The more general question put aside of whether two sentences, uttered by different people, can have the same meaning (Section 6.4), it suffices here to recall that pronouns (*she*) or deictic expressions (*that guy*, *here*) are considered to have the same meaning although they may point to different real world entities depending on the context in which they are uttered. *She* for example can be said to have a stable meaning because the word always points to the most easily available female person in the discourse.

However, we did not manage to obtain satisfying results, neither with statistical nor with prosody-

based tools and the sentence segmentation was done manually, even for the large Orféo treebank. Although the notion seems quite well-defined, even the inter-annotator agreement on the manual segmentation into illocutionary units does not seem to be very high, but we have not conducted any systematic study on the question in our annotation projects. Yet, the application of the same segmentation rules to corpora that students created as part of my classes on corpus linguistics revealed important difficulties to agree on sentence segmentation and gave rise to extensive discussions.

This may sound worse than it actually is. If the segmentation is too large and contains two independent illocutionary units, the roots of these units will simply be connected by a link that represents the paradigmatic relation between the two units.[40] And it would not be difficult to cut up units that contain two paradigmatically linked configurations that fulfill certain criteria, for example the criterion to be composed of two finite verbs that each have their own subject: The sequence *This afternoon, Mary sleeps and Peter watches TV* may be separated into two illocutionary units by these criteria even if the initial cutting-up still combined them, independently of the scope of *this afternoon*. As long as the parser does not choke with the length of the sentence, it is better to miss some sentence-final punctuation than to put in one too many. For sentence segmentation we prefer false negatives to false positives.

This also goes to show that the whole debate of sentence length is actually artificial in the sense that we can always connect two subsequent units to a larger one by adding a conjunctive link. Any text can be seen as one big list of conjunctions of sentences, possibly with embedded structures, which gives a tree structure resembling what is done in Rhetorical Structure Theory (Thompson & Mann 1987). And inversely and more importantly for our discussion of linguistic infinity in Section 6, many sentences can be split up into smaller units. This illustrates that including coordination or other paradigmatic relations in the list of dependency relations actually mixes apples and oranges. Coordination is more of a discourse relation than a syntactic relation.[41] If we only consider pure syntagmatic relations we easily reach the upper limit of embeddings (see Section 4.1.1) where a reformulation is possible and seems advisable, at the very least for stylistic reasons.

5.5.2 Lower limits

The lower limit of syntax, i.e. the units that we tend to combine by our syntactic structures are the words. Contrarily to Chinese with its *scriptua continua* free of spaces, this notion might seem quite obvious for written English and French: Words are the letters between spaces. That works most of

40 In the Universal Dependencies annotation scheme (UD, Nivre et al. 2016, annotation guide on http://universaldependencies.org/u/dep/index.html), the *conj* relation is used for constituent coordination with or without the presence of a coordinating conjunction, whereas sentence coordination only uses *conj* if the sentences are connected by a coordinating conjunction and *parataxis* if such a conjunction is absent. In **Gerdes & Kahane 2017** we propose to use a dependency name that relates these two types of conjunctions: *parataxis:conj*.

41 Of course this does not hold for all coordinations. For example *Peter and Mary wear a hat* can be paraphrased by *Peter wears a hat. And Mary wears a hat.* But *Peter and Mary carry a piano* cannot be paraphrased by *Peter carries a piano and Mary carries a piano*, yet, the sentence can be understood in two parts something along the lines of *Peter and Mary form a group. This group carries a piano.* This question of collective and distributive readings of the conjunction *and* has been an important subject in formal semantics, in particular in the 1990. The first part of the previous paraphrase can be seen as the introduction of a “collectivity operator” as proposed in Winter (1996).

the time quite well although there are well-known difficulties ranging from contractions (*don't, du* Fr. 'of the') to hyphenation (*check-in, est-il* Fr 'is he'*),* idiomatic expressions (*one size fits all, la main dans le sac* Fr 'red-handed'), and named entities (*Bertrand Russell, Pays de la Loire*). We have developed recommendations for the dependency annotation in general in **Gerdes & Kahane (2016)** and annotations of these problematic cases in particular in **Kahane et al. (2018)**. We advocate for syntactic annotations inside idiomatic constructions as long as the regularities are visible to a native speaker of the language. This is the case for the French *pomme de terre* 'potato' which is recognizable as an "earth apple" to a French native using the common construction *noun de noun*. Equally, idioms, consisting by definitions of more than one token, are in general syntactically compositional. Special multi-word annotation should only be used in the rare cases where the relation between the words cannot be recognized such as in English, for the city name *Hong Kong*, which, however, could be analyzed as just a common adjective-noun relation with some knowledge of Cantonese where it means 'perfumed harbor'. Things get a bit tricky here because this recommendation makes annotation depend on the language knowledge of the annotator: Any speaker of French would recognize the words *de la* in *Rio de la Plata* even without any knowledge of Spanish and any German speaker recognizes the *van* in *Ludwig van Beethoven* as a variant of *von* 'from'. Would a Spanish speaker recognize the Arabic *al* in *Al Jazeera* as a determiner similar to the Spanish *el* 'the'? The problem is actually of a more general nature: Annotation always depends on the domain knowledge of the annotator as it is hard to annotate a research paper in chemistry without knowledge of the subject matter, or even an article about movies is hard to parse without knowledge of authors and movie titles.[42] We shall return to the size of a language's vocabulary in Section 6.3 .

5.6 Paradigmatic piles

Most linguists say that spoken data is more central to the analysis of language than written data, which is too strongly influenced by normative prescriptions and dependent on years of training (Bloomfield 1933, Gleason 1955, Blanche-Benveniste 2010). Yet, spoken language has never been the center of any sub-science of linguistics with the obvious exception of phonology and phonetics. In syntax, even the most basic notions such as constituents are frequently defined by means of prosody (Osborne 2018) but then the syntacticians do not actually assess the specificity of spoken language. We claimed "No syntax without prosody" (**Gerdes & Kahane 2000**) but kept looking for simple formalism whose simplicity is tested on strings of texts and on implementations trying to generate or analyze written data, leaving the phonetization to a separate module. For us just as for many other syntacticians, the hurdles of spoken texts were too high and the gains appear too distant or even uncertain as long as we wanted to describe language competence.

Although it is only a few years ago, retrospectively, it seems astonishing that the Rhapsodie project

42 The editorial of *Les Cahiers du Cinéma* can be hard to read as in this example "La Pointe courte est pensé sur une alternance comme Les Palmiers sauvages de Faulkner, Sans toit ni loi est construit autour de douze travellings qui raccordent secrètement sur la musique de Joanna Bruzdowicz, Le Bonheur, à la colorimétrie extraordinaire, est l'un des plus grands films de peintre qui soient." (June 2018) 'La Pointe courte is conceived on an alternation like Faulkner's Les Palmiers sauvages, Sans toit ni loi is built around twelve tracking shots that secretly connect to Joanna Bruzdowicz's music, Le Bonheur, with its extraordinary colorimetry, is one of the greatest painter films ever made.'

(**Lacheret et al. 2014**) that we started in 2008 was actually the first project for any language to systematically annotate spoken data with both dependency and prosodic annotation. Prosodic annotation aside, it was also the first project developing a syntactic treebank of spoken French, and *a fortiori* the first project to systematically annotate natural French spoken data with a dependency annotation. There was just one other comparable project that we know of, the Corpus Gesprooken Nederlands (Corpus of Spoken Dutch, CGN, Schuurman et al. 2004), but they explicitly skipped disfluencies, hesitations, and reformulations, and their dependency links actually connect a selection of the transcribed words that form reconstructed "simple" sentences following the usual written conventions.

Yet, these disfluencies, hesitations, and reformulations are the first feature that catches the eye when looking at transcriptions of real-world spoken language data. So our take on the spoken data was heavily influenced by the only people we had heard of who actually do syntax on spoken data: the Aix School, that had developed what they call "macrosyntax" on spoken data (Blanche-Benveniste et al. 1979, 1990). They had developed the notion of "list" that combines segments playing the same role in the utterance, and which can normally be replaced by one another. We formalized this notion in combination with a dependency analysis under the term *paradigmatic pile* (**Gerdes & Kahane 2008**). The idea is that constructions such as disfluencies (*I I I I really want that*), reduplicative intensifications (*very very very good*), hesitations (*I am a linguist, uh, maybe a mathematician*), and reformulations (*I am a linguist, a language scientist*) share with coordinations (*I am a linguist and a mathematician*) that they actually describe paradigmatic structures. It is even oftentimes difficult to tell these constructions apart, in particular for spoken data. So when initiating the corpus annotation, by pure pragmatic necessity, we started to give these paradigmatic constructions the same analysis, leaving it for a later more fine-grained annotation to actually make the distinctions. It turned out that it is indeed possible to establish a hierarchical typology of paradigmatic phenomena, but the distinctions are hard to make and the fine-grained annotations are not easily reproducible (the inter-annotator agreement is not satisfactory, Kahane & Pietrandrea 2012).

5.7 Learning to understand

Christianini (2014) shows that *knowledge* is a declining term in computer science whereas *learning* is on the rise. It is very telling that the new scientific paradigm is machine *learning*, not machine *understanding* or machine *knowledge*. What is happening is that we are learning without gaining knowledge. The problem with any scientific paradigm in a Kuhnian sense is that a new tool puts us in a new mindset, we are hit by Maslow's hammer: If all you have is a hammer, everything looks like a nail. Christianini goes on to say that "scientists are very good at restricting the universe of their discourse to what they can formalise. When they stop seeing the rest of the world, they might convince themselves that they can formalise everything."

The previous scientific paradigm was the deduction from a set of axioms. The job of the researcher was to find those axioms. Reasoning was symbolic manipulation. The need for hand-encoding of increasing amounts of symbolic knowledge as exemplified by the Cyc project (trying to build an ontology of everything, Lenat et al. 1989) showed the limits of this approach: Humans find it hard

to keep control over such large amounts of data, making maintenance and improvements a near-impossible task, similarly to what happened to other rule-based linguistic systems in Machine Translation or grammar construction, such as the impossibility to construct a real-world TAG meta-grammar reported in Section 4.1.2 .

Historically, we can say that linguistics as a science started out as a collection of languages, to be classified into a language tree. Then it moved to a rather short-lived “global” theory in structural linguistics, particularly generative syntax, which was not particularly empirical and non-operational for NLP. Now linguistics is moving back to stamp collection, its original paradigm, by embracing treebank construction.

It is tempting to think that future linguistics will consist only of a sheer description of the phenomena, while finding the over-arching rules, optimized combination and exclusion of rules can be left to the machine. Yet, it is not possible to even start an annotation of language data without a first classification. A transcription of spoken data is a classification task. Recognizing a part of speech or a syntactic function is another classification task. And some first version of classes must pre-exist to these first steps. Machines can help in this classification task: Unsupervised clustering algorithms can propose different clusterizations that have to be validated by the annotator. And, inversely, different annotations can be checked for their (supervised) machine-learnability, which is one kind of internal coherence of the annotation.

What does it actually mean that an annotation is easy to learn for a machine? If annotation *A* contains less information than annotation *B*, it is certainly easier to learn, the most extreme case being that *A* assigns the same value to every word. If however *A* and *B* are isomorphic in the sense that one annotation can be transformed into the other by lossless and simple rules, then, any difference in learnability would rather measure the limitations of the machine-learning algorithm – the limitation to learn precisely those simple transformation rules from the given training set. I do not know of any “difficulty” measures other than the error rate itself to assess how difficult the actual learning process was for the parser. But if the error rate is consistently higher for annotation *A* over *B* over a range of different parsers, we may conclude that *B* actually is the “better” annotation scheme. This is the case for example in a function-word-headed dependency structure, which gives better parse results than a content-word-headed analysis (Schwartz et al. 2012, Silveira and Manning 2015, Kirilin and Versley 2015, Rehbein et al. 2017). We use this argument to advocate the Surface-syntactic annotation scheme SUD over the near-isomorphic existing UD structure (**Gerdes et al. 2018**).

In another previous work (**Guibon et al. 2014 and 2015**), we similarly tried to use machine learning tools to actually understand language data. The research started out as a straightforward practical problem: We have some small treebanks for Old French and some more raw data of Old French that we would like to annotate. The data is highly heterogeneous because it stems from a large period of a few centuries, and moreover, Old French has not undergone any important normalization process. The simple question was: What treebank data to use for training a syntactic dependency parser in order to analyze each of the texts that remained to be annotated. Should the training data be from the same region as the text to be parsed? Or is it more important that the

training data stems from the same period? Or is the style that makes the difference?

These questions are not completely new. They commonly appear in what is known as “out-of-domain parsing” (Nivre et al. 2007) and “cross-language parser adaptation” (Zeman & Resnik 2008). Apart from the small training and evaluation data size as well as the particularities of the non-normativity of the data, we gave an important twist to the interpretation of the test training results: The error analysis of the parser allowed us to map the proximity of the different texts, by vocabulary and syntactic constructions. The idea is simply that similar data is good training data. The degree of learnability of the one annotated text from another annotated text shows the proximity of the texts. The black box of statistical parsers becomes a torch for understanding historical language data.

Surely, this is still a primitive utilization of machine learning tools for actual linguistic understanding. Maybe the extractions of desirable linguistic rules can be learned directly, too, if enough training data is provided, just like beautiful faces can be detected – and most importantly subsequently generated? What we would need is a way of representing many example rules that we would accept as pleasant generalization. Previous feature-engineered statistical parsers had thousands of rules, most of which seem completely meaningless or even wrong to a syntactician – it’s their weighed interplay that makes the parser functional. We would have to simplify, group these rules to make them interpretable. Maybe machine learning algorithms could be trained to firstly distinguish pleasant from useless rules, and then, in the next step, be used for the discovery of rules that have similar properties of being acceptable and useful for humans.

The simplification of complexity is actually what all machine learning is about. For example, the most famous neural machine learning tool for NLP is Word2vec, which is an effort to reduce the large dimensional space of word co-occurrences into fewer (at most a few hundred) dimensions in a way that the topology implies semantic or syntactic proximity. This allows for the famous computation *king - man + woman = queen*. It is not exaggerated to say that all recent progress in parsing and Machine Translation can be attributed to these tools because they allow actually making use of raw unannotated data, which was very difficult for the previous feature-engineered parser generation, a problem known under the term of “self-training” (Yu et al. 2015).

These tools are in constant evolution, the most recent offspring being ELMo (“Embeddings from Language Models”), which performs much better on polysemic words (Peters et al. 2018). The clue in these systems is that the dimensional reduction is not only necessary to adapt to the limits of current computer power and corpus size, but also represents some kind of smart generalization over the data.

In view of these developments, it seems less absurd to expect machine-learning tools to actually produce simplifications of the data of adjustable degree of complexity; simplifications that can be represented as interesting, satisfactory rules. But is it true that “*from all this it should be possible in the next decade or two to crack the semiotic code, in the sense of coming fully to understand the relationship between observed instances of language behaviour and the underlying system of language*” (Halliday 2005) and the prediction came just a little early? What kind of linguistic rules can realistically be expected? Will the new rules really be communicable and satisfactory to the

linguist's mind? Or do we run the risk of receiving answers that are true but are not more meaningful than Douglas Adams' 42? Maybe data-driven linguistics is just like a self-driving car: There is no reasonable doubt about its perfection but it does not fit in well with human behavior and expectations.

5.8 Embracing language complexity

It is clear that the current machine learning algorithms are trained to be the best possible predictors: Predicting the next word, the best translation, the best tag for a word. And following Hockett's assumption that the linguist's task is prediction, linguists should just sit on the sideline to watch Machine Learning improve, with better algorithms and more data. But is it correct that understanding language is the same thing as being able to predict? What does it actually mean to generalize over the data that one has seen?

Interestingly, the difficulty of generalization is the central problem of machine learning, it is the aforementioned problem of *overfitting*. In a sense, the machine learns by heart the whole list of provided data. The neural network is optimized for exactly all the data it has seen. But nothing else.

To prevent or at least reduce overfitting in neural network, a so-called regularization technique is used that goes under the name *Dropout* (Srivastava et al. 2014). The method seems quite brute: It simply consists in ignoring, "dropping out", i.e. not updating the weights of randomly selected neurons during training. This results in more robust representations because it requires the network to construct multiple representations of the same connections, the same rules. And this robustness is better prepared for unseen data with unseen words and constructions.

In some way it seems necessary to be redundant to make better predictions. Simplicity has to be engineered away explicitly to get good results. Do we have to learn from this also in view of the linguistic rules we should expect?

In an influential paper (Halevy et al. 2009), researchers at Google advised us linguists to stop hoping for simple formulas. While accusing economists of suffering from "physics envy", they recommend to all science describing human behavior that "*we should stop acting as if our goal is to author extremely elegant theories, and instead embrace complexity and make use of the best ally we have: the unreasonable effectiveness of data*." And the data they prone is not human annotated data but rather data that is out in the wild: raw, unannotated data.

Should we give up our search for regularities? Maybe not quite yet. It seems necessary though to adjust the scope of the rules we want to discover.

We see that the empirically best system is redundant. More specifically, it may be true that even if rules can be found that are empirically true, they may not be the optimal way of capturing what is actually going on in the language. Moreover, it is unlikely that they have a neural reality in the human brain because of the randomness of the neural training procedure. Even without random weights, slight changes in the neural network setting give very different resulting network structures. So it seems unlikely that we are actually producing anything like a model of the brain when training a neural network.

Winograd's supposedly modest remark that AI is at the stage of alchemy "pouring together different combinations of substances and seeing what happens, not yet having developed satisfactory theories." (Dreyfus, 1992) – a quote that Thierry Poibeau puts at the very beginning of his book (2011) – is still filled with this profound hope of redemption from complexity. The hope that behind all that empiric chaos lie the lush fields of logic.

Yes, linguistics still has to look for simple rules. Yet, we need to transform what we consider simple. Simplicity will no more be the simplicity of the system but rather the simplicity of the usage of the tool. This can be seen as another step in the ongoing outsourcing of human knowledge: Firstly we outsourced some memory functions by the invention and optimization of writing systems towards books, printed books, calculators, electronic storage, shared electronic storage (internet), smart access (search engines). Now we are on the brink of storing the model itself in the machine.

A simple system can then mean two things: In a predictive sense, it means an optimization of the system's capacity to handle seen, and more importantly, unseen data. Such a system probably stores syntax, semantics, and world knowledge somewhere in its neural structure, but it has no explanatory power for us. In an explanatory sense, we need some downgrading, and a simple system produces rules sufficiently simple so that they give us the pleasure of (the illusion of) understanding and communicating about the rules.

While the Renaissance and the industrialization developed a technology that made us see the infinity of the universe, the 21st century's technology makes us see the limitations of human cognition.

The fact that the best predictive language rules can be machine-learned from a moderately small corpus – just some chunks from the Web, a far cry from infinity – also reminds us of the limited creativity of Language. We gladly succumb to the illusion of freedom in Language: We want to believe that Language allows us to say everything we want to say. In reality, languages tend to oblige us to say things we do not actually intend to say (see **Kahane & Gerdes in preparation**, chapter 1.2) such as the Japanese obligation to mention the status relation that we perceive with the interlocutor. Grammar "*determines those aspects of each experience that must be expressed. When we say, 'The man killed the bull,' we understand that a definite single man in the past killed a definite single bull. We cannot express this experience in such a way that we remain in doubt whether a definite or indefinite person or bull, one or more persons or bulls, the present or past time, are meant. We have to choose between these aspects, and one or the other must be chosen*" (Boas 1938, p. 132). This mainly West European distinction between definite and indefinite determiners actually means that we have to express whether any object we mention can actually be uniquely identified by the interlocutor. In all these cases, the native speaker is so used to this distinction that it takes the view from another language to actually perceive the incongruity of such rules. "Languages differ essentially in what they *must* convey and not in what they *may* convey" (Jakobson 1959). These general rules imbue our thinking and do not have much of any beautiful universal nature. Moreover, for any language, there are loads of idiosyncratic lexical rules to learn, namely idioms, but also fixed sentences that are more or less idiomatic, such as the way to say that the sun shines in the morning, discussed in Section 4.2.2 . Speaking a language well means to know

all those idiomatic structures. It means to master the repetition of what others have said before.

The next chapter will explore what this means for innovative language use and more generally for human creativity.

6 On the exhaustion of language

Up to here I have discussed why historically and epistemologically, linguistics have been looking for simple rules while at the same time being obsessed with the infinity of language. I have shown that it is more likely that rules are better if they are learned from large corpora or treebanks, and yet these rules are too complex for a human mind. It is likely that there is nothing better waiting for us behind the green mountains, although we can shrink the rules to fit our mental capacities for teaching or scientific pleasure.

In this chapter, I want to discuss why language is not infinite and what that means for linguistics, for our notion of creativity, and for the understanding of ourselves. To make my reasoning easier to follow, I will now give an overview of my line of thought.

We start by asking what kind of infinity we are talking about, actual infinity or cognitive processing incapacity in a lifetime. We then discuss that we can limit our considerations to sentences and do not need to consider longer language utterances because illocutionary units have a meaning independent of context, which is sufficient to show our point. If we now see Language as the Saussurian dichotomy of the signifier and signified, we have two ways to tackle language infinity: by showing the finiteness of the number of sentences or of the number of thoughts.

Starting with the former, we have to come back to the lower and the upper units of sentences. By a quite lengthy discussion that allows me to return to the problem of transcribing spoken data, the starting point of my treebank creation experience, we will show that the transcription can be seen as text annotation with a very large tagset: the vocabulary. And Language also possesses a systematic way of presenting and including new words. The answer is less clear-cut on the possibility of having sentences of unrestricted length. Yet, even if we assume that there can only be a large but finite number of illocutionary units, it is still possible that there is an infinite number of ideas – and some ideas just do not have means to be expressed by language. This point will have to be conceded as long as we consider that ideas can exist without language in a continuous space of perception. But as soon as we restrict ourselves to “useful” ideas, i.e. “ideas worth spreading”, we discretize and reduce the space further by applying a grammar as well as a limited vocabulary.

We then go on a legal excursion to the world of patents, defining terms such as *claim*, *novelty*, and *inventive step*, and discussing the consequence of enumerability and finiteness of ideas for the area of patents. Then I will finally describe the most astonishing project I have been involved in, namely the construction of an actual Library of Babel: patent claim variation in Cloem. Before concluding, I will briefly turn to the generalizability of this approach to other domains of intellectual property and to the notion of artificial creativity.

6.1 Infinity and the like

10^{17} seconds have passed since the universe exists. That is a large number. But the Traveling Salesman Problem, the most famous problem in optimization, asking how to find the shortest path visiting a set of cities on a map has a whooping 10^{655} solutions for 318 points (Moscato 1989). That

is a much larger number. Both of these numbers transcend our cognitive capacities when thought of as a list or a sequence of events and one might be tempted to equate actual infinity and numbers like this. Yet, there are two reasons why the question matters whether a language contains an actual infinity of sentences or not: Firstly, smart algorithms can find the optimal path for the Traveling Salesman Problem in spite of the astronomical number of solutions. And for any algorithmic approach, it is important to know the size of the space to be explored in advance. For language, too, smart algorithms might be able to make smart predictions if we describe well the available space. For example a Natural Language Generation (NLG) algorithm might be conceived differently knowing that we have to solve a choice problem rather than actual generation. This is in keeping with the second reason: Awareness of the finiteness of linguistic choice makes it clear that each sentence that has been pronounced for the first time has not been created but chosen from a list of possible sentences. The act of pronouncing the sentence for the first time reduces the space for other speakers of the language to produce a "new" sentence. This gives a new role to a central notion of today's society: Creativity. We will discuss this in the last Section 6.6 of this chapter.

Language is the most familiar and everyday point of contact between the human mind and very large numbers. It is very hard to grasp that these large numbers are not infinity, and when Émile Borel (1913) attempts to illustrate the proof of the Law of Large Numbers, he used the famous thought experiment of one million monkeys typing on a million typewriters who would eventually generate work by Shakespeare. Borel uses the example in order to give "une idée de l'extrême rareté des cas exceptionnels, rareté qui dépasse tout ce que notre imagination peut concevoir." 'an idea of the extreme rarity of exceptional cases, a rarity that exceeds anything our imagination can conceive of'. Yet, the Law of Large Numbers holds true and the monkeys will really type Shakespeare at one point, given enough time. In particular, we can update the illustration a bit and replace typewriters by powerful computers and monkeys by computer algorithms that have access to vocabulary lists and grammar rules.

6.2 Honey, I shrank the Language

In Section 5.5.1 on the upper limit of syntax, we have introduced the notion of *illocutionary unit*, akin to a sentence, but being defined not by punctuation but by the illocutionary unit's stability of meaning independent of context. As illocutionary units are segments of thoughts, we may limit ourselves to illocutionary units when discussing linguistic creativity. Complex thoughts that cannot be expressed in one sentence may exist but the identification and definition of this type of complexity can only be done once the simple, one-sentence, thoughts are well understood. We leave that question to further research. At least in the patent domain, one-sentence thoughts are seen as sufficient for the description of everything that is worth privatizing, as we will see in Section 6.5.1 .

We will start with the signifier of sentences, and look at the maximal length of written sentences. I have argued that actual illocutionary units are commonly short if we consider sentence coordination to be a discourse relation between, but not inside, sentences. We have already shown that center embeddings are difficult to process, and just three of them bring any sentence to the brink of agrammaticality (Section 4.1.1). We will explain in Section 6.3.2 that in these cases it makes sense to posit an interpretability threshold and not to use induction in the attempt to show that it is

always possible to go one step deeper into incomprehensibility. It remains to be discussed whether language also restricts other forms of embeddings, in particular right branching structures, also called *edge embedding* in opposition to center embedding. Edge embeddings are clearly easier to process and can go much deeper than center embeddings. Take sentence (2) as an example.

(2) *Yesterday, the guy came to see me to whom the lady talked that day that I met you in the bar where Beth celebrated her wedding with the cousin of the guy from the badminton class, where I go every Monday except when you ...*

Even if such sentences with edge embeddings are clearly better than sentences with an equivalent number of center embeddings, they nevertheless remain very bad style. This means that they are hard to process and that alternatives exist – paraphrases that cut this sentence in different sentences according to the different thoughts expressed. An example of paraphrasing (2) is given in (3).

(3) *Yesterday, a guy came to see me. The lady talked that day to that guy. I met you in that bar the same day. Beth celebrated her wedding in that bar. She married the cousin of the guy from the badminton class. I go every Monday to that class. I do that except when you...*

The problem with (3) is not each sentence's processing complexity. The reader does not know where the story is going because the sentences in (3) lack a coherent discourse structure, i.e. words that help to find the unifying thread, the point of what the speaker wants to say.

We hypothesize that edge embeddings also have an upper limit in the sense that complex ideas can and must be cut up in smaller chunks of thought. The whole text may get slightly longer and less elegant when edge embedding are resolved into multiple juxtaposed or coordinated sentences by means of pronouns or deictic structures, just like translations may get longer if the target language does not dispose of the same lexical unit or grammatical constructions but the meaning can still be rendered in another language. We hypothesize further that this possibility of resolution of complex edge embedding holds for any human language because the same cognitive constraints make it necessary to allow for simple thoughts to be held in memory. Infinite recursion does not exist in language because there is a limit to sentence complexity. In other words, we believe that there is no thought that requires more than say 10 embeddings of any kind to be expressed, and that there is no language that has a sentence with 10 embeddings that the natives consider necessary or good style. The actual number can be derived from treebank measures for different languages. This idea brings us back to the epistemological question of what we consider good answers in Science: An explanation that comprises 10 or more independent elements that cannot be grouped or categorized in any way seems unsatisfactory because it is not simple enough. – It may have missed a generalization. Our cognitive memory and processing limits equally shape grammar and science.

Even if this conjecture is wrong, for what follows, it is sufficient that most interesting sentences have an upper limit of embeddings. If we stumble upon a counterexample, the theory could be adapted and updated (or not, as discussed in Section 3.3.2). Now that we have put a bound to syntax, it remains to verify that the complexity does not hide somewhere in the vocabulary.

6.3 Infinite words?

"We have reason to believe," continued the lama imperturbably, "that all such names can be written with not more than nine letters in an alphabet we have devised." – "I see. You've been starting at AAAAAAA . . . and working up to ZZZZZZZZ. . . ." – "Exactly – though we use a special alphabet of our own. Modifying the electromatic typewriters to deal with this is, of course, trivial. A rather more interesting problem is that of devising suitable circuits to eliminate ridiculous combinations. For example, no letter must occur more than three times in succession." – "Three? Surely you mean two." – "Three is correct: I am afraid it would take too long to explain why, even if you understood our language." […] – "Well, they believe that when they have listed all His names – and they reckon that there are about nine billion of them – God's purpose will be achieved. The human race will have finished what it was created to do, and there won't be any point in carrying on. Indeed, the very idea is something like blasphemy." – "Then what do they expect us to do? Commit suicide?" – "There's no need for that. When the list's completed, God steps in and simply winds things up . . . bingo!"

This extract from Arthur Clarke's novel "The Nine Billion Names Of God" not only illustrates that the interest in infinity transcends the world of linguists. It gives the basic ideas of what we have to discuss in this section: Can we devise an alphabet? Do we have reason to believe that all words can be written with a number of letters below an upper limit? Can we invent a new name (for God)? Does the establishment of a language's vocabulary boil down to a simple question of combinatorics? And what happens once we have all the words?

While the finiteness of the vocabulary is only implicit in Humboldt's thoughts (Section 3.1), the idea that not only the list of phonemes but also the number of words of a language is limited precedes Chomsky by at least two decades. Boas (1938) already stated without any further justification that "the number of words that is, of groups of sound symbols used in any one language for various aspects of experiences cannot be unlimited." This undoubtedly depends on what we call a word, and we have started discussing the notion of *word* in Section 5.5.2 on the lower limits of syntax. We will now discuss this question in greater detail by recounting the treebank creation process for spoken language.

6.3.1 Transcription

A spoken-language specific problem of the definition of words is the transcription itself. We will discuss now that a transcription can be seen as a categorization of a section of sound with the entries of a dictionary. Learning how to write is actually nothing but recognizing sound bites and linking them to a written symbol (a sequence of letters or an ideogram). It is an interesting question to ask whether this dictionary is finite or infinite – or even uncountable with continuous, non discrete components – because supposing finiteness, countability, or uncountability results in different views on where complexity can be found in Language and how to account for it. We will show that the question does not have a clear answer that emerges from the linguistic analysis. Yet it is possible to make the vocabulary finite, depending on what the linguist actually considers and annotates as different words.

It may seem trivial to assess that different people can use the same word, the same sign. On the one hand, this means that language users have a similar, but not equal, signified in the sense that it refers to different sets of referents, for example people would disagree on whether a certain object is a chair or not. We can say that the extension of a sign varies. On the other hand, they can also refer to the same sign although pronouncing the signifier differently, variants known as allophones or, if spelled differently, allomorphs. We will return to slight variations that express the exact or nearly the same meaning in Section 6.5.3 , but then on the level of sentences: Sentential allomorphs.

An interesting debate we had during the beginning of the Rhapsodie annotation project was whether *oui* 'yes' and *ouais* are allomorphs. Diachronically, that is certainly not true as the Wiktionary explains: "Before becoming an informal synonym of *oui*, the word was exclamatory. According to Auguste Scheler, it could be linked to Indo-European roots: *οὐαί* (ouaí) in Greek, *vae* in Latin, *uau* in Portuguese, *wow* in English, *wau* in German, *guau* in Spanish or *vai* in Romanian" (https://en.wiktionary.org/wiki/ouais). Today, in addition to the confirmation, *ouais* can have an ironic or skeptical undertone (Peroz 2009) more akin to *yeah* or *yep* in English. Slight distributional differences between the two words tipped the scale in favor of the distinction of *oui* et *ouais*: It is possible to say *"Je pense que oui"* 'I think so, yes' but not "[??]*Je pense que ouais"* and "si oui..." 'if so' but not *"*si ouais"*. Yet, these distributional differences are also related to the language register: The casual tone of *ouais* may be incompatible with complex reasoning as in "*Je pense que oui*" and "*si oui..."*.

Similarly, *ouioui*, *ouiouioui*, *ouiouiouioui* can be seen as paradigmatic piles of *oui*. Yet, they do not seem to intensify as *very very* does, rather the contrary. So *ouioui* is a different word from *oui*. But is *ouioui* different from *ouiouioui*? Likewise, the interjection *ouh là* 'watch it' is a warning, adding more "là" as in the all-French *ouh là là* 'wow' 'jeez', makes the expression more of an admiration. Again, it is unclear whether *ouh là là* is different from *ouh là là là*.

To give an example directly from English, in the 11 million word corpus of rap lyrics (collected by my Master student Madeleine Hernandez, Hernandez 2018), we find 94 variants of *yeah*:

> *Y-E-A, ye-ah, Ye-ah, yea, Yea, Yea, YEA, yeaaaaahhhh, YEAAAAAHHHH, YEAAAAH, yeaaaah, Yeaaaah, Yeaaaahhhhh, YEAAAAHHHHHH, yeaaaahhhhhh, Yeaaaahhhhhhhhhh, yeaahhhh, yeah, yeah, yeah, YEAH, Yeah, yEAH, yeah-eah, Yeah-eah, yeah-eah-eah, Yeah-eah-eah, yeah-eahh, yeah-eee, yeah-hah, yeah-hea, YEAH-HEAA, yeah-heh, Yeahh-hah-hahhhh, yeahha, yeahhh-ha-ha, Yeahhh-haaa, yeahhh-hahhh, yeahhhAHHHHahhhh, YEAHHHH, Yeahhhh, yeahhhh, yeahhhh-eh-eh-eh-eahhh, Yeahhhh-heahhhhh, YEAHHHH-YEAHHHH, yeahhhhh, YEAHHHHH, Yeahhhhh, Yeahhhhhh, YEAHhhhhh, yeahhhhhh, YEAHHHHHH, YEAHHHHHHH, yeahhhhhhh, Yeahhhhhhh, Yeahhhhhhhh, YEAHHHHHHHH, yeahhhhhhhh, yeahhhhhhhhh, YEAHHHHHHHHH, Yeahhhhhhhhh, Yeahhhhhhhhhh, YEAHHHHHHHHHH, YEAHHHHHHHHHH, Yeahhhhhhhhhh, yeahhhhhhhhhh, Yeahhhhhhhhhhhh, yeahhhhhhhhhhhh, Yeahhhhhhhhhhhhh, yeahhhhhhhhhhhhh, yeahhuh, yeahhuhh, YeahI, yeahI, yeeaaa, Yeeaaaah, Yeeaaah, Yeeaaahhh, yeeaaahhhhhhhh, yeeaah, Yeeaah, yeeah, Yeeah, YEEAH, Yeeahh, Yeeea, yeeea, Yeeeaah, yeeeah, YEEEAH, Yeeeah, yeeeahhh, Yeha, yeha, yyyeah.*

Not to mention the duplicate *yeahs*: *Yeahyeah, yeahyeah, YEAHYEAH, Yeahyeahyeah, YeahYeahYeah, yeahh-yeahh, yeahhyeah, Yeahhyeahhh,* and the similar word *YAY, yay, Yay, YAYY, yayyy.* Did the poets refer to the same sign? Was it just screaming louder when they transcribed with all upper case? Or do some variants appear in very different contexts, triggering very different connotations, so that it would be wrong to group them together?

The answers to all these questions seem difficult to generalize. It depends on which degree of detail we want to achieve in our transcription, whether we expect this (register?) difference to be relevant for the study we want to make subsequently on the corpus. In this point, we see that we face the same question as when developing a tagset, for example for the POS annotation. We cannot start tagging a corpus without fixing the distinction that we need, and we cannot even transcribe a text either.

We can see the problem as a juridical question that could be central in a court question: If the two words are sufficiently similar to denote the same meaning, it should not be possible that "He said *oui*" is true and "He said *ouais*" is false in the sense that he agreed. Although purely logically, the two sentences do not have the same truth condition – one can pronounce the confirmation more closely to *ouais* or more closely to *oui* – one could conclude that these words have the same *juridical truth condition*. We will come back to this question in Section 6.5.3 .

When designing a corpus that requires a transcription, it is quintessential for a homogeneous transcription, not depending on the transcriber, that these three problematic cases are well described:

1. How to deal with non-standard pronunciations for which exist spelling variants. For example *wanna* vs. *want to*, or *chépa* vs. *je ne sais pas* 'I don't know'. 2. How to deal with inaudible words and creative onomatopoeia, just imitating sound. And 3. How to deal with words from foreign languages and other appearance of reported sound without meaning for the addressee. These last points have been used as an argument for the infinity of the lexicon.

> The foreigner screamed neeeeeeeeeeeeeeeeeeeeeeeeeeeee.…
>
> The cow went 'moo'/'mooo'/'moooo'/'mooooo'/'moooooo…
>
> The space alien said 'klaatu barrada nikto' to Gort. (oratio recta)
>
> My car goes 'ehhrgh' when I go over a bump. (example from Pullum & Scholz 2001)

Postal 2004 gives the above examples to conclude that "there is no viable reason to believe that while e.g. English permits reporting as in [the first example] human expressions or as in [the second example] cow noises of various lengths, at some point there is a human or cow noise too long to be reported. What rule of English fixes a maximum length on reportable noises? Since it is impossible to specify an actual bound, the answer must be none. In short, and for several reasons, it is then impossible for every piece of direct speech […] to be mentioned in the grammar of any [Natural Language]." The aforementioned start of Chomsky's Syntactic Structures (Section 3.3.1), shows the importance given to this finiteness of the vocabulary in the generative framework: "From now on I will consider a language to be a set (finite or infinite) of sentences, each finite in length and

constructed out of a finite set of elements." The stipulation that the vocabulary is finite is maintained in more recent texts by Chomsky such as Chomsky (2000). To me, it is actually not clear why this finiteness holds that primordial importance to generativists, other than that it upholds the Humboldtian *infinite use of finite means*. Even with an infinite vocabulary, the same grammar generating the supposedly infinite set of sentences could be developed. Of course, a hypothetical infinity of the vocabulary would call for some type of generative lexicon, that structures the infinite list of words. It would make morphology just as central in the discovery of the miracles of Language as syntax. And it would be closer to what Humboldt had in mind because he "takes the word to be the basic unit of analysis [whereas] Chomsky's focus is the sentence" (Losonsky 1999, p. xxxii).

6.3.2 How to welcome new words

It is clear that Language allows introducing new terms. "A faculty of speaking a given language implies a faculty of talking about this language. Such a 'metalinguistic' permits revision and redefinition of the vocabulary used." (Jakobson 1959) During their lifetime, humans will likely encounter new lexical units among the open classes of parts of speech such as nouns, verbs, adjectives, and adverbs – and no new prepositions, pronouns, or determiners. And Language is prepared to add these new encounters to its lexicon. The introduction of new terms can be done explicitly as in these two sentences:

> *This new procedure will be called **xyz***
> *This new procedure will be called **zyx***

Or it can be done by simply using the terms in a context where the reader can learn the usage of the word, and thus has access to the meaning in a Wittgensteinian sense. The most famous example illustrating this dynamic 'live' introduction to new terms is probably Nadsat, the language of the novel "A Clockwork Orange" by Anthony Burgess. One of the first sentences of the novel reads: "Our pockets were full of **deng**, so there was no real need from the point of view of **crasting** any more **pretty polly** to **tolchock** some old **veck** in an alley and **viddy** him swim in his blood while we counted the takings and divided by four, nor to do the ultra-violent on some shivering **starry** grey-haired **ptitsa** in a shop and go **smecking** off with the till's guts. But, as they say, money isn't everything."[43]

The bold words are not English, although some of them use the English inflectional system. The reader can guess for example that *deng* is the Nadsat term for money, which is simply the Russian word *деньги den'gi* 'money', because in the 1950s, when Burgess wrote his novel, it seemed reasonable to imagine the future of language as a mixture of the two dominating forces of the time.

Schtroumpf, the fictional Smurf language, functions similarly: "The fundamental rule of Schtroumpf is: Replace every term of the common language with *schtroumpf* as often as you can

43 The bold face was added by me. My personal experience with this novel is that I received a copy from an American friend of mine when I was 17, and I abandoned the reading after a few pages, writing back to her that my English is not good enough for this text and that even my dictionary does not include these words. She wrote back to me explaining Burgess' idea of Nadsat. The second attempt worked out better, and I still believe that this novel is an important piece of world literature for its style but also for the interesting thoughts on Ethics it triggers.

without excessive ambiguity." "The Schtroumpf language is a parasitic language, because, although nouns, verbs, and adjectives are replaced with the all-purpose homonym, it would not be understood if it were not backed up by the syntax (and the various lexical contributions) of the base language (be it the original French or its translations)." (Eco 2000, p. 278) Note that both Nadsat and Schtroumpf rely on the high redundancy of Language which makes it possible for speakers to be understood even if they omit many lexical units. Information theoretically, we can say that Language has been conceived for very noisy channels. This high redundancy may seem surprising if Language's function is seen mainly in the transmission of information, omitting its ritual, group building role that may be even more fundamental.

Yet, Nadsat is different from Schtroumpf because the latter only introduces one additional highly homonymous term: *Schtroumpf*. Nadsat actually presents new lexical units, borrowed from Russian and modeled after the English connotational space. One way to accommodate for these additional words is to consider the dictionary to be instantiated while parsing the text. So, when *xyz* and *deng* is added to the dictionary, it still remains finite. This requires some kind of two-step parsing process where the unrecognized sequence is first handled as such, and then the syntactic and semantic analysis of the sentence checks whether we have to revise the dictionary.

If the answer is "yes, we do have to revise the dictionary", we parse again with the new dictionary that now includes the new term, either as defined in the sentence itself or just with the information of the present context: Burgess' sentences allows us to know that pockets can be full of *deng* – which is already a valuable syntactic information (*deng* is a non-countable noun) and semantic information (*deng* is a physical substance that can fill pockets). We already have an incomplete dictionary entry.

Another important aspect of these sequences is that we cannot transcribe anything but the phonemes we know. A good example is the recent surge of language pundits insisting that China's capital is no longer *Peking*, the designation that had been used for centuries, stemming originally from the Cantonese pronunciation[44] of the city, but rather *Beijing*. When asked, these people insist that this is how the Chinese say. Leaving aside the fact that they do not use *Roma* 'Rome', *Praha* 'Prague', or *Zhongguo* 'China' either, the funny thing is that they chose a transcription in their own phonetic system. They forgot that the Chinese do not say *Beijing* but rather *Běijīng*, with a third tone on the first syllable and a first tone on the second syllable. *Beijing* is not like the Chinese say, but it is the closest transcription of how the Chinese say in the English phonetic system – and if a Chinese pronounces *Běijīng*, all they hear is *Beijing*, but they will also hear *Beijing* when a Chinese says *bèijǐng* 'background'.[45]

What is important to take home here is that even if space aliens or cars talk to us as in the above

44 The Cantonese pronunciation is also closer to the Old Chinese phonology which would have pronounced something like /pəkˉ kʲiæŋ/ – "would have" because the city did not bear that name yet. But the pronunciation used in all European languages for the city most probably stems from the contact between Europeans and Cantonese sailors.

45 Similar to European languages and their Greek-Latin-based mostly technical vocabulary, Japanese borrowed and keeps creating words, called *kango*, made up of Chinese morphemes deprived of their tones. This results in a significant proportion of homophones in Japanese, which does not seem to hinder communication in Japanese – because of the aforementioned redundancy of human language. I have tried, yet without any success, to convince Chinese people to give up their tone system based on the *kango* experiment.

examples, depending on our musicality we may be able to reproduce the sound more or less faithfully, but we can only write, and possibly think, the sound within our phonetic system. Regardless of the phonetic rules allowing only specific sound combinations, there is an upper bound for the possible combinations of phonemes: The 44 phonemes of English (or maybe just 35? See Bizzocchi 2017) cannot form more than 44^6 sequences of length 6. /beɪˈd͡ʒiŋ/ is one of them, and /b*ěijīng/* is not. So when we encounter unknown sequences while transcribing spoken language, we first have to ask whether the sequence is a sound that can be transcribed phonetically.

If the answer is no but the sequence plays a syntactic or semantic role that we want to include, we have to find a meta-language for filling this gap, either just indicating the non-transcribable gap as in "The deaf person went ___ yesterday" or by describing the gap as in "[Piano student plays passage in manner μ]. Teacher: 'It's not [plays passage in manner μ]. – It's [plays same passage in manner μ']" or in "Sheila went ___ but I did not make that gesture." (examples from Postal 2004).

If, however, the sequence is not the signifier of a sign of the text's language, but we can transcribe the sequence phonetically, we can either use the phonetic system of the text's main language,[46] possibly indicating in our transcription that this "word" is not part of our vocabulary and requires special treatment when parsing. Or we just use the indication of a gap, an unknown word at this position. This second, simpler procedure would not make a difference for the parsing process of a single sentence – the computer (or the human parser) has to take educated guesses on the nature of the word, depending on the context in which it appeared. The parser has to handle each word of the vocabulary and an additional token UNKNOWN, triggering special rules. However, not transcribing the word in any way, would not allow any dynamic extension of the system's vocabulary and would not be suitable for tasks such as the detection and analysis of neologisms.

Depending on the required subsequent linguistic analysis, we can thus 1. restrict our vocabulary to a finite set plus a special token for unknown words and gestures, 2. allow for the dynamic addition of new transcribable tokens, or 3. start out with the idea of a potentially infinite vocabulary made up of phonemic combinations, whose combinatorial explosion can be limited by syllabic rules and, more importantly, a maximum number of phonemes per word, keeping the vocabulary finite.

From the standpoint of transcription, it is clearly possible to allow for a finite number of phonemes per word. Put simply, there is no need to allow infinite letters to transcribe the cows' mooooooo sounds – a few "o" letters suffice to show the extreme length of the utterance. And if we have reasons to believe that the utterance is structured, it is probably possible to cut up the unknown sequence into smaller words, just as we proposed for coordinated sentences in Section 5.5 .

Of course, any time we fix a limit, say 50 phonemes, it is unclear why 51 phonemes should not be allowed as a word. This is how the inductive argument goes. Yet, in nature there are many phenomena of fuzzy borders. If we say that the day ends at 8:50 P.M. it seems strange to assert that 8:51 P.M. is in the night. If, however, we choose the day's ending time way into the dark, the actual time may be arbitrary, but it still constitutes a clear maximum of what can be considered a day.

In Language as well, if we take Postal 2004's example "My father's father's father's father's father

46 Note that bilingual speakers who speak language *A* actually have to learn to pronounce words from language *B* in the phonetic system of language *A* if they want to be understood by a native of *A*.

died", we can say that this sentence is already too far off into indecipherability to reasonably assert that it is not possible to add even another "father's".

It has always astonished me how easily structuralist grammarians and their successors, in particular Generative grammarians, put a strict border on the continuum of grammatical to ungrammatical sentences, labeling one sentence grammatical and a very similar variant as ungrammatical, independently of frequency or comprehensibility of the sentence (see Section 3.3.1 in particular Footnote 23). They could just as easily put aside sentences and new word creations that go beyond a certain length. But this is completely inconceivable to them – because this step would imply the finiteness of language.

6.4 Discreteness

We have seen that the same signified can have many signifiers (the variants of *yeah*) and that the same signifier can have a whole set of signifieds (*schtroumpf*). In most cases the correspondency is one to one. Also, we can now exclude that there is an infinity of meaningful sentence signifiers because the syntactic complexity of illocutionary units as well as the vocabulary can be given an upper limit. But can there be an infinity of different signified, of different thoughts? There clearly is a continuous space of perception. Yet, it is unclear whether it makes any sense to say that feelings and impressions can be thought about before they become conscious, as in Wittgenstein's (1933–1934) famous discussion of "unconscious toothache". It is quite certain, however, that feelings and impressions can be conscious without being put into language, evidence of which is that humans regularly encounter the situation of having a perception while not being able to put the right word or expression to it in order to share the perception. At the same time, many interior thoughts clearly take place in form of an interior discourse, as a set of sentences in a language. So it may well be that some meaningful feelings or perceptions do not have a linguistic equivalent.

I hold the basic positivist belief that an actual thought can be expressed (or at least approximated) and thus shared linguistically, that ideas can be transmitted, that memes exist (Section 2.3). This seems like a trivial assumption but it goes against fashionable identity science where it is assumed that the personal experience determines thought and, thus, for example the outcome of scientific research. Yet, nobody would seriously defend that there are 7 billion versions of Fermat's theorem, in the sense that the meaning actually differs. (There are certainly more than 7 billion accoustic variants, i.e. allophones, and possibly 7 billion allomorphes when doing all the combinatorics on synonyms and structural variations such as diathesis changes, see Section 6.5.2). Holding the belief that ideas can be shared amounts to nothing less than believing in our capacity of empathy, the aptitude to mentally put oneself in the place of another person, not perfectly, but well enough to be compassionate. Doubting this empathetic capacity is a political statement that threatens the social fabric of our society (Sokal & Bricmont 1998). This does not discredit the notion of connotation: A cat meme can trigger different feelings in different people depending on their prior life. And to some extent these feelings can be shared linguistically.

Boas (1938, pp. 125 – 126) discusses the impossible linguistic uniqueness of every experience : "*In gestures as well as in spoken language large numbers of situations are expressed by the same*

formal expression. In other words, a grouping together of similar experiences persisted, and these have found their representation in symbols. Thus dog, water, running, sleeping, expressed either by gestures or by sound, include each large numbers of varied experiences, all understood under a single vague symbol. Other aspects of the same experiences may be expressed in similar ways. By means of the combination of these symbols, the whole situation is described from various angles and is thus communicated." In a sense, the belief that ideas can be shared is a belief in the capacity of discretization, it is Bloomfield's (1933, p. 78) fundamental assumption of linguistics: "we must assume that in every speech community some utterances are alike in form and meaning" and *discreteness* is one of Hockett's (1955) design features of Language.

Prosody (as well as the gestures that Boas mentions) has discrete features called *prosodemes*, but it allows for maaaany ways of transmitting continuous perceptions. As we have argued in the preceding section, we want to consider these forms as different pronunciations of the same word, we want to group yeeeahs and moooos. This grouping can be seen as the concept of finite fields in mathematics[47], which are, in fact, a factorization of continuous space into a finite number of elements. Each element of the finite field groups together an (even uncountable) infinity of points in a continuous field. And each vocable groups together an infinity of perceptions, of neural states.

Note that discreteness does not imply finiteness. There could be a countable infinity of semantic units if some word forms had an infinite list of meanings. In the Meaning-Text Model (Mel'čuk 1988), for example, there is a discrete set of meanings and a discrete set of (spoken or written) texts. To my knowledge, no claims have been made on the finite or infinite size of these sets. This allows me to mention sloths one more time by way of a quote from Bloomfield (1933, p 145): *"The quality sloth and the animal sloth probably represent a pair of homonyms to some speakers and a single meaning to others. All this shows, of course, that our basic assumption is true only within limits, even though its general truth is presupposed not only in linguistic study, but by all our actual use of language."* But even if there should be a word form with a countable and infinite list of meanings, all these meanings are discrete and could be triggered by one word-form. Once again this seems implausible mainly for cognitive reasons: A lexicographer who described a vocable in this way may be seen as having missed some important generalization of the vocable's meaning.

Just like for individual word forms, we may assume that a sentence covers a segment of continuous semantic space, i.e. the sentence describes an infinity of states of the world. And two sentences may be sufficiently similar so that some of that space is covered by both sentences. Thus, it seems conceivable to completely cover a domain of infinite continuous situations by a finite set of sentences.

6.5 Privatizing semantic space

The notion of Intellectual Property (IP) includes a whole set of legal configurations such as

47 Finite fields are particularly interesting if the number of elements is a prime number. These "prime fields" have come to fame because they constitute the space in which we do elliptic-curve cryptography. It is easier to multiply prime numbers than to find the decomposition of a number into its prime parts. This can be translated into an encryption where it is harder to decrypt than to encrypt, the very basis of Public Key Cryptography. Without public keys, we could not have secure web pages with an httpS. This was one of my favorite subjects in algorithmic geometry and I have followed fascinating classes on the matter in Utrecht and Paris.

copyright (© first: 14 years, now: 50 to 100 years), trademarks (™ ® 5 to 10 years, renewable), patents (20 years), design rights (13 to 25 years), and trade dress (aesthetic models, ~ 3 years). All forms of IP have in common that some semantic space is attributed to one (juridical or legal) person for exclusive use. Even copyright, which originally was only concerned with the exact form of a text, increasingly has a semantic component as it now includes derivative work such as movies with the same or a similar story as the book. The existence of IP rights is justified partly by the danger of confusion for the consumer of goods (in particular for trademarks), but mainly by the concept of creating "incentives", i.e. the belief that only with the legal protection of ideas ensuring return on investment enough people would want to be creative and invent new ideas and technology, which in turn benefits the whole society. Alternative concepts, for example around the Free-culture movement and Copyleft as well as Open Access groups, have put forward that creativity is always a form of remix whose very existence is conditioned on pre-existing work (so-called "derivative works") and a context that is favorable to "creation". IP laws, on the contrary, are seen as hindering creativity. Activists from the free and open-source software community were mainly worried about patents as "patent law […] poses one of the most significant threats to the open code movement that there is" (Lessig 2002). This threat was foremost seen in software patents, which are essentially the protection of an algorithm that, contrarily to copyright, hinders a reprogramming of a software with the same functionality.[48] For example, the steps of a procedure where one entity informs the other of stock value changes by calling an API may be patented. The Electronic Frontier Foundation, opposing patents, attributes a monthly negative price to the "Stupid Patent of the Month" where even more absurd real world example are commemorated.[49] Some of these patents read "a little like what might result if you ate a dictionary filled with buzzwords and drank a bottle of tequila" writes EFF lawyer Daniel Nazer.[50] Yet, even "strong" patents, as considered by patent professionals themselves, result from mere linguistic manipulations. And creating a computer system that does just that, eating a buzzword dictionary and mixing it up, is a humorous but not inaccurate way to describe the goal of the Cloem project that will be presented in Section 6.5.2 .

It is hard to prove or disprove one or the other point of view on patents and other parts of IP, as this would require a social experiment on a global scale. But, the question of whether patents should exist may be obsolete because, as we shall see, the whole field can be expected to undergo profound changes triggered by the emerging technology of creative machines.

6.5.1 Patents

Among political activists, from the open-source movement and beyond, patents have no good press: "*The old dinosaur notions of pinning down ideas with patents and so on will be seen clearly for what they are — an attempt by individual egos to colonize inspiration that ultimately inhibits*

48 More recently, patents and open-source software coexist more or less peacefully because problems have shifted to the distribution of code and its usage on the Cloud where patented methods may remain undetectable (as in complex or hardened hardware, where reverse engineering can hardly be performed). So-called "software patents" are increasingly mainstream and the scope of protection encompasses a functionality rather than verbatim copy considered by copyright.

49 https://www.eff.org/issues/stupid-patent-month These cases can also be seen as an excess of a bureaucracy (since many of those patents ended up being revoked or highly amended).

50 https://www.eff.org/deeplinks/2014/11/stupid-patent-month-who-wants-buy-teamwork-penn-state

human thriving. Just as we learn to stop trusting the individual ego in others to save us, we will stop trusting it in ourselves, and increasingly we will hand over our internal control to the greater part of our consciousness that does not seek to control, define, claim and deflect, but rather draws from a wisdom that often eludes narration but creates harmoniously." (Caitlin Johnstone 2018[51])

The exponential growth of patent applications as well as of granted patents in the last years can be explained by the growing importance of ideas compared to physical objects but also by systemic reasons: Any administration strives to grow and patent granting organizations have no incentive to reduce the number of patents as each patent filing funds the organization. Legal changes that would strengthen the requirements for new patents, and thus reduce the number of patents, have been proposed, for example under the Obama administration in 2014, but "large, patent-rich technology companies such as Microsoft and IBM [...] lobbied hard against a proposal to expand a patent office program to invalidate more low-quality software patents" (Lee 2014). The natural move for Patent Offices thereby supports increasing compartmentalization of inventions as the concept of transposition is absent from patent laws. For example, a patent could be granted for ancient Egyptian hydraulic principles as soon as it is used in microfluidics. This opens wide the gate for automation of such domain transpositions in patents.

Jurisdictions in the world have different but similar, if not converging, criteria for examining and granting patents. In Europe, the conditions comprise *novelty* (understood as *non-existing), inventive step* (to be discussed below), and *industrial applicability* (increasingly downgraded to a secondary criterion). The equivalent concepts in the USA are *novelty*, *nonobviousness*, and *usefulness*.

No automatized tests exist for these requirements. They all have to be interpreted by humans that have to invoke a fictional legal person called "the skilled person", numerous Case Law decisions, Supreme Court decisions, and of course the guidelines for examination themselves. This may lead to a high number of litigations invalidating previously granted patents (Abbott et al. 2018). "For decades, obviousness has been the most common issue in litigation to invalidate a patent, and the most common grounds for a finding of patent invalidity" (Abbott 2017).

A patent application has different parts, such as an introduction, the technical field (for classification purposes), a description, drawings, and patent claims, each of these parts having a very different syntactic and lexical structure (Oostdijk et al. 2010). Claims, the legal core of the patent, defining the boundaries of the patent's legal protection, have a particularly codified structure that needs to be understood for parsing them. One claim is usually a single extremely long noun phrase (*A method of making a stabilized mayonnaise spread, said method comprising the step of incorporating an effective amount…*), sometimes preceded by a fixed phrase such as "*We claim …*". Many claims contain lists such as

… by a process comprising the steps of
(i) inoculating a pasteurized dairy composition with ... ;
(ii) incubating the composition until ... ; and
(iii) separating the whey from …[52]

51 https://medium.com/@caityjohnstone/why-do-our-heroes-always-let-us-down-5e258fdc02b0
52 Example from patent US6113954A. Not all lists are spread over multiple lines.

Secondary claims can specify and thus become dependent on a previous claim:

2. The method of making the stabilized mayonnaise spread described in claim 1 wherein the fat content of the mayonnaise is about 33% or less.

The raison d'être of dependent claims is precisely to keep the more general first claim as broad as possible and, at the same time, prevent a more specific follow-up patent that would claim the additional details compared to the first claim.[53] Depending on the jurisdiction, the number of dependent claims and multiple dependencies can be restricted because of their important interpretative combinatorics.

The novelty criterion requires that an idea in order to be patented has not been disclosed (i.e. implemented or exposed) anywhere before, that there is no collision with *prior art*. "Prior art is any evidence that your invention is already known."[54] So any publication describing an idea makes the idea unpatentable.[55] This leads to the untenable situation where "[t]he person of ordinary skill [for purposes of determining obviousness] is a hypothetical person who is presumed to be aware of all the pertinent prior art" (Hattenbach & Glucoft 2015, citing from an actual patent litigation case). This "skilled person" has neither been objectified nor embodied, and Wikipedia represents the closest representation to date. The free prior art databases Espacenet, operated by the European Patent Office, itself contains 90 million documents, let alone all other technically relevant texts in any library or in the internet. Submitting a patent application has thus become a bit of gambling that a relevant document of prior art does not exist or will not be found. As discussed in the previous Section 6.4 , the textual content contained in new patent applications and newly published texts further reduce the remaining "space of invention".

Technological solutions to the patent inflation have been proposed, such as the European PATExpert project lead by Leo Wanner (2006) in cooperation with the European Patent Office (EPO). One of the ideas of the project was to propose an ontology of elements to be used by the patent applicant when describing the new idea, in particular in order to expose the novelty in relation to existing patents. The technological challenges put aside, such a system had no chance of actually being put into service because it neglects the interest of the two parties involved in a patent application: The goal of a good patent attorney is precisely not to be too clear about what is to be claimed in order to protect a wide field of unforeseen future developments or actual implementations of the idea. In a

53 A "follow-up" or "improvement" patent is a patent describing a specification or slight modification of an existing patent. This can be granted as long as the novelty of the more specific idea is above a "threshold of originality", i.e. the more specific idea has described a sufficiently *nonobvious inventive step* with respect to the existing patent. A frequent result of such a patent distribution is a mutual dependency. A simplified example might make the situation clearer: If you obtain a patent on a bicycle, an adverse party, having read your published patent files, gets a patent on "a cycle with a bell". You are now somehow blocked in your commercial developments by the second patent (you cannot sell your bicycle with a bell), while the adverse patent needs a license from you to sell the bike (with or without a bell). So-called *patent thickets* can be inextricable: A smartphone today is said to correspond to tens if not hundreds of thousands of patents (CPU, network, UI, etc). This contributes to the formation of oligopolies with patent litigation non-agression pacts. A startup, producing a novel type of phone, can be overrun with litigation trials and pushed out of business (or forced to sell out).

54 From the EPO's Inventors' Handbook: https://www.epo.org/learning-events/materials/inventors-handbook/novelty/prior-art.html

55 This is true at least in Europe. In the US the rules are a bit more complicated as the authors of a disclosing text themselves can still file a patent for a limited time.

dialog with the patent examiner, the patent attorney tries to place as many general terms as possible into the text in order to give wider semantic scope to the text, but also tries to give examples and hyponyms of the general words in order to optimize the scope of protection, creating “intermediate generalizations” in legal speak, and thus to prevent the aforementioned more specific follow-up patents. The patent examiners, as a representative of the general public, strives on the contrary to reduce the semantic scope of the patent so that the first applicant is not given too much legal power that would privatize a too wide monopoly in the semantic space of inventions. This can be seen as a language game, where both participants play with synonyms, antonyms, hypernymes, hyponyms, and meronyms, in order to claim as much semantic territory as possible, nearly independently of any planned or actually existing invention.[56] Each word has an impact on the size of the semantic footprint of the claim. A slight variation of a claim, a near-paraphrase, might be seen as the same meaning in a litigation case, or it might not be. That is why another proposal of the PATExpert project, to provide simplified paraphrases of patent claims in order to give easier access to the massive amount of documents (Bouayad-Agha et al. 2009), was also doomed. It is a bit like proposing automatic paraphrases to a philosopher of Plato. The philosopher will not be interested because every word of the holy text counts.

6.5.2 Cloem

Several legal loopholes in patent laws, amplified by developments in technology are leaving the door wide open to the assault of smart machines. Machines are already used to explore a well delineated field of inventions, for instance the best aerodynamic shape of a new Japanese bullet train (Plotkin 2009). “This new approach to problem solving is the cause of a number of legal and ethical issues. Today, machine learning may be used to create computer-generated patent claims; this computer-generated content has a wide variety of potential applications that could wreak havoc on the patent legal system as it currently stands. Examples of machine learning in patent law include composing computer-generated claims to be utilized by companies to generate prior art to invalidate potential infringing patents, or by inventors to improve their claim language or actual inventions” (McLaughlin 2018). Ryan Abbott points out that “an inventive machine standard will dynamically raise the current benchmark for patentability … [and] it may be that every invention will one day be obvious to commonly used computers. That would mean no more patents would be issued without some radical change to current patentability criteria” (Abbott 2017), but, as we have seen, the profiteers of the current status quo have a strong lobby holding on to the current laws and preventing any quick legal change.

One of the projects exploring and exploiting legal loopholes is Cloem, a French legaltech startup from Cannes, for which I worked as a scientific adviser during my sabbatical years from 2011 to 2013. Their name is a portmanteau word for *claim poems*, which describes their project of building an actual Oulipo machine, a program for potential literature, English technical poems in the patent domain. The user provides a so-called base claim and chooses the technical domain (the patent classification code, thus guiding the vocabulary selection), and Cloem analyzes the claim syntactically, identifying terms, expressions, part of speech, and modifying optional clausal

56 This is not true for the domain of chemical and medical inventions, where studies have to be provided for the actual effectiveness of the proposed process.

elements. Terms and expressions are systematically replaced by synonyms, hypernyms, and hyponyms, while modifiers are optionally suppressed. For the following claim, for instance, many words and expressions will be replaced, among them the two words in bold face:

A method comprising the steps of:
creating a hash value for a prior ***transaction****;*
combining *the hash value, transaction data and the public key of a transaction recipient; ...*

The word “transaction”, to give a simplified example, has the paradigm of synonyms *transaction, payment, share, exchange, swap, compensation, purchase* and “combining” has the paradigm *combining, joining, authenticating, reinforcing, blocking, transforming, releasing*. This can be illustrated like a context menu that opens at each word or expression, and any combination can be chosen as in Illustration 6.

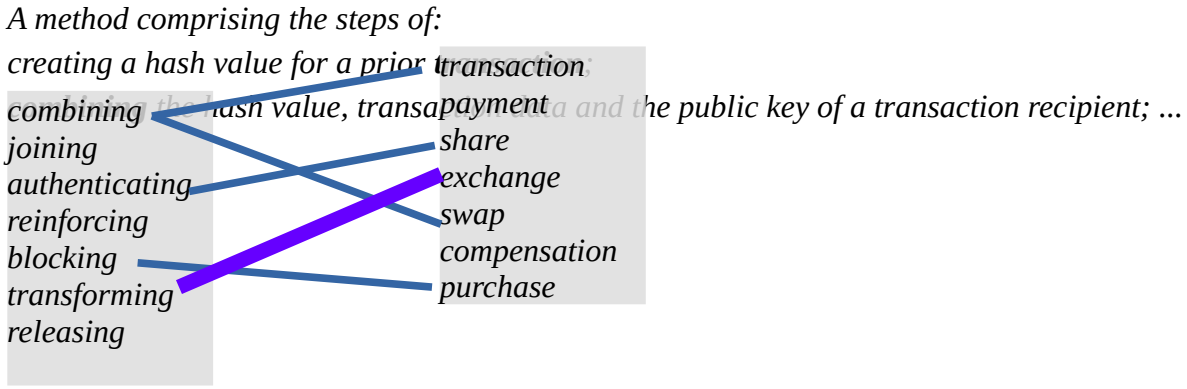

Illustration 6: Near-synonym paradigm choice on a claim text depicted as context menus.

The thick purple line of Illustration 6, for example, corresponds to the claim “*A method comprising the steps of: creating a hash value for a prior* ***exchange****;* ***transforming*** *the hash value, transaction data and the public key of a transaction recipient; ...”*

In this way, for each base claim, billions of similar claim texts are created, timestamped, and stored in a database where they are freely searchable. Cloem is allowed to publish texts with very small variations and even the original text because published patent applications are exempted from copyright. Each text has a slightly different semantic scope and is associated with a unique URL address. The replacing words are selected with the goal to thoroughly cover a space around the base claim. To avoid the generation of texts describing methods that the user might want to patent later or simply keep private, the user has complete control over the words used in the Cloem texts thus avoiding all unwanted cross-pollination or self-collision.

Cloem is a neutral technology that can serve multiple causes, in a pro- or anti-patent perspective. For the common good, Cloem can serve as a public domain operator, for example by covering the semantic space of entire technical domains, thus avoiding any patents as a published idea cannot be patented. Cloem can also be used for private purposes, to prevent improvement patents (these patents are filed after having read a pioneering invention and can lead to situations of mutual dependence). Many intermediate situations also become possible, combining defensive publication and the search for private rights.

Other organizations, such as Allpriorart (http://allpriorart.com), attempt similar effects on the patentable domain by means of Massive Defensive Publishing. Contrarily to Cloem, Allpriorart creates texts by training the language model on existing patent applications, non-systematic reshuffling, and publication of a much smaller set of modified patent claims. Existing organizations such as IP.com just publish one unique version of a defensive publication, at high prices and without any robustness against adverse variations.

6.5.3 Consequences of Massive Defensive Publishing

Massive Defensive Publishing opens a whole new domain of Natural Language Processing that is quite different from existing domains while some techniques can be reused. For examples, the central task of Natural Language Generation is to find the optimal textual realization of a meaning representation, because only that one text will be shown to the user. In Massive Defensive Publishing, however, we want to generate a very large, but finite, set of sentences that together cover a semantic space. We need to be equipped with domain specific lattices of near-synonyms of technical terms, a challenging need in itself as many words and expressions are brand new – patents do contain novelty after all, also on the lexical level. My former Master (and current PhD) student Li Yixuan has worked on this technology for Chinese technological texts (Li 2018). It is planned that this type of resource development will be a focal point of my research delegation at the Inria team Almanach starting at the end of 2018.

The goal of covering a semantic space is contingent of knowing how to estimate the distance between two texts. Are two sentences close enough to describe the same state of affairs, are they sentential allomorphs? Based on the complete lattice of near-synonyms we might be able to estimate the number and size of steps one has to take on the lattice to transfer one sentence into the other, a metric akin to the aforementioned Levenshtein distance. Replacing perfect synonyms and changing grammatical voice (active to passive voice for example) should result in a zero distance. Other sentence transformations, lexical or grammatical, have to be ranked according to their impact on the semantic space described by the resulting sentence.

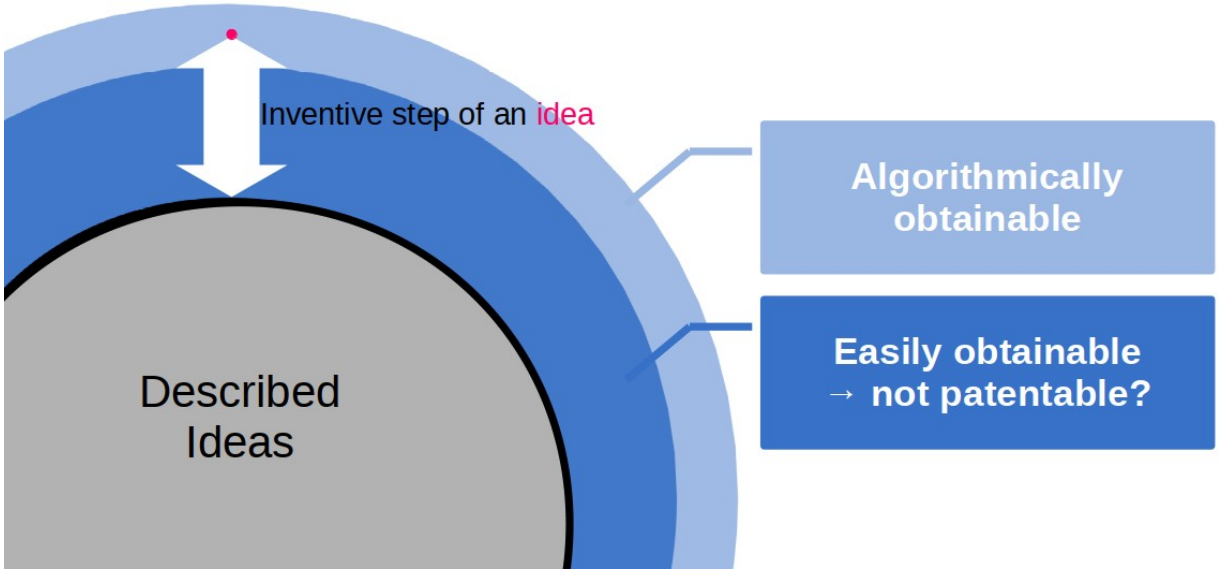

Illustration 7: The distance of an inventive step.

Evaluation of Massive Defensive Publishing algorithms is difficult and different from other branches of Natural Language Generation. One way could be to present stochastic samples of the generated Cloem texts to patent attorneys in order to assess the quality. Technical and financial difficulties of this approach put aside, the question remains whether the patent attorneys themselves can actually evaluate the future usefulness of a text in a potential patent litigation if that evaluation is not based on the analysis of large amounts of textual patent data in order to detect trends. In the Conclusion we will return to the question of predictability of future technology.

An actual "proof" of Cloem will have to wait until an actual patent litigation case will consider a Cloem text as "enabling prior art". The developers of Cloem have developed fallbacks in the event of an evolving Case Law (e.g. papers' publications of samples of generated series, increasing role of humans versus the inventorship question, etc). The questions raised by the project are unseen and possible consequences on the legal system are difficult to estimate as of today. Yet, it seems likely at this stage that the inclusion of machine-generated prior-art in the patent law will result in a steep reduction or even complete elimination of what can be considered patentable.

Not everyone welcomes such a development. McLaughlin (2018) warns that "*in addition to computer-assisted inventions, where humans utilize sophisticated software to assist in design, there are other ways machine learning can rattle the current patent landscape. One example of this exists in the generation of prior art references. A French company, Cloem, is using "algorithmic patenting" to computer-generate claim language. The company operates by taking existing patent claims (or an invention description if a patent is unavailable), and creates thousands of timestamped linguistic variants of the text called cloems. While this particular use of machine learning technology may benefit a singular inventor by defending him against other potential patentees who may seek to circumvent the existing invention, it is harmful to society as a whole. Such iterations create prior art, not necessarily directly covering the submitted invention, that may prevent other potential inventors from obtaining legitimate patents; inventions that would be independently patentable. As a result, said inventor would likely not invest in developing his invention covered by an iterative claim because he is not guaranteed the protection he deserves, and the invention covered by the iteration could in effect sit unused by society.*"

Naturally, I do not share this prediction of the negative impact of Cloem. There have been numerous reports of the contrary: Ideas that cannot be put into production because of threatening patents. Moreover, Cloem, at least with its current technology, does not directly represent a threat to inventions with an important inventive step. Yet a continuous flooding strategy based on existing base claims and clever vocabulary choices, imitating the impact of inventors reading journals such as Wired and Technology Review in order to catch inventions that are "up in the air", could wipe out numerous weak applications.

And it is questionable whether patent claims that have been obtained by simple lexical replacement and grammatical transformations constitute "legitimate patents". I would predict that Cloem-like technology will eventually be used directly in the evaluation process of the patent offices: A Cloem distance measure of the difficulty of obtaining the submitted application from the existing patent

texts may provide an objective measure and set a standard of the application's inventiveness.

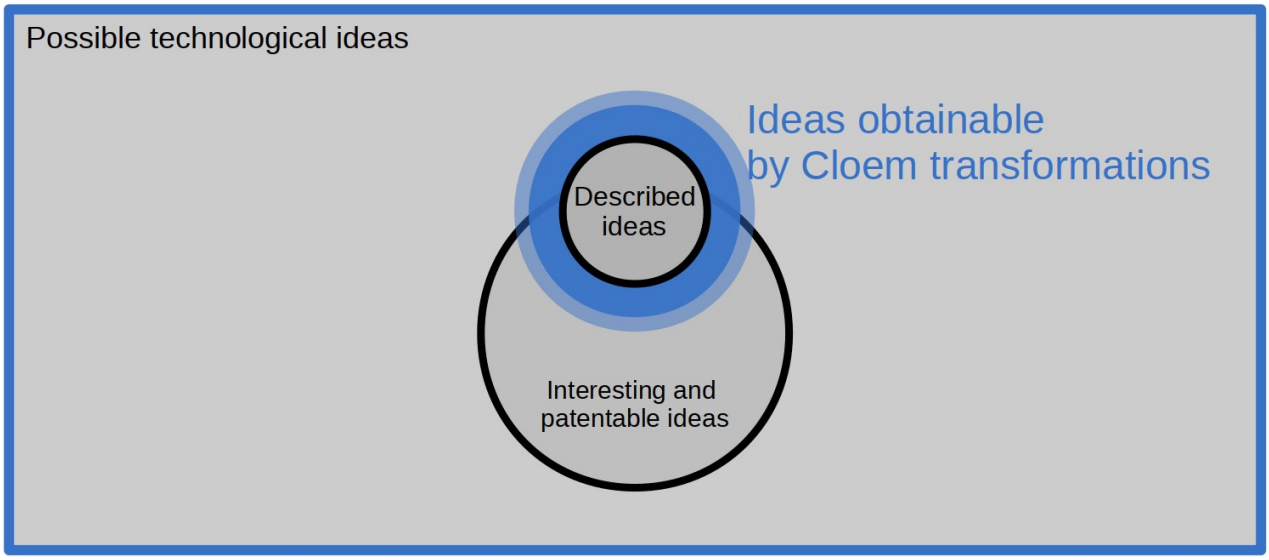

Illustration 8: Ideas obtainable by Cloem transformations.

Note that Massive Defensive Publishing à la Cloem does not actually generate complete ideas; only one side of the linguistic sign, the signifier, is actually generated. The signified, a meaning representation only evolves when a human reads the text. Not that this would be completely impossible to do, we could in parallel generate a semantic graph for each Cloem text. But just like for human-generated texts, it is possible that the human reader's apophenia, see Section 2.3 , creates new links, i.e. adds unintended information – unintended in the sense that the human creators themselves did not see that link in the text and that the Cloem variation algorithm added an unforeseen idea.

Frege's old debate from the Philosophy of Language pops up here: People who do not know that the evening star and the morning star both refer to Venus, would not know that it is contradictory to state to have seen the morning star but never the evening star. It is thus interesting to replace X even by its perfect synonym Y because the meaning may be different in intensional contexts, i.e. when the context requires a mentioned person to know the equivalence between X and Y. For instance *A method to produce X* may have a different intensional meaning from *A method to produce Y* if the audience does not identify the synonymy relation between X and Y. So the *juridical truth condition* introduced in Section 6.3.1 is essential in the patent domain, too: Two patent claims C_1 and C_2 that have the same juridical truth conditions should be considered equal. This means here that one cannot claim C_1 without also claiming C_2, which depends on the semantic width that is given to a sentence, i.e. the semantic space that the claim covers, the breadth of actual real world configuration of ideas that are considered described by the claim.

6.6 Creativity and the finitude of language

Si las páginas de este libro consienten algún verso feliz, perdóneme el lector la descortesía de haberlo usurpado yo, previamente. Nuestras nadas poco difieren; es trivial y fortuita la

circunstancia de que seas tú el lector de estos ejercicios, y yo su redactor.[57]

It is interesting to point out that the approach of Massive Publishing cannot easily be applied to other fields of Intellectual Property. It is clearly possible to take a copyrighted text and apply Cloem-like transformations but it would not be possible to publish these texts if we do not hold the copyright on the original text ourselves. Moreover, as Hattenbach & Glucoft (2015) note, "*the [US] Copyright Office has already announced that it will not register works produced by a machine or mere mechanical process that operates randomly or automatically without any creative input or intervention from a human author.*"

Can our notion of creativity subsist in a world where we choose from a large but finite set of ideas and sentences? That certainly depends on how we define creativity.

Firstly, in a trivial sense, it is clear that the absence of creativity, in the narrow sense of an actual unpredictable creative act, is a straightforward result of the absence of free will. Yet, independently of a theoretical predictability on a neural level, we commonly consider artists creative if they transgress pre-established rules, if they move the goal posts. Such rule-changing creativity can be opposed to rule-governed creativity. When Chomsky (1962) states that "a theory of language that neglects [the] creative part of language is only of marginal interest", he refers to rule-governed creativity. For Chomsky, following rules is creative, reminiscent of the creativity of the "art académique", the 'academic art', which is often opposed to actual creativity and demeaned as "art pompier" 'firemen art'. Scientists who are rule-governed creative remain in Foucault's (1966) *grid of thought*, inside Kuhn's (1962) paradigms, where science only consists in exploring the complete expressible space. A scientific paradigm is the set of thinkable and expressible ideas including possible questions, answers, and interpretations. It is the syntax of science. Interestingly, Kuhn does not ask the question of the finiteness of the expressible space. In any case, "real" creative scientists are supposed to overcome dogma and bias and strive to trigger paradigm shifts.

57 'If the pages of this book contain some successful verse, the reader must excuse me the discourtesy of having usurped it first. Our nothingness differs little; it is a trivial and chance circumstance that you should be the reader of these exercises and I their author' (Borges 1923).

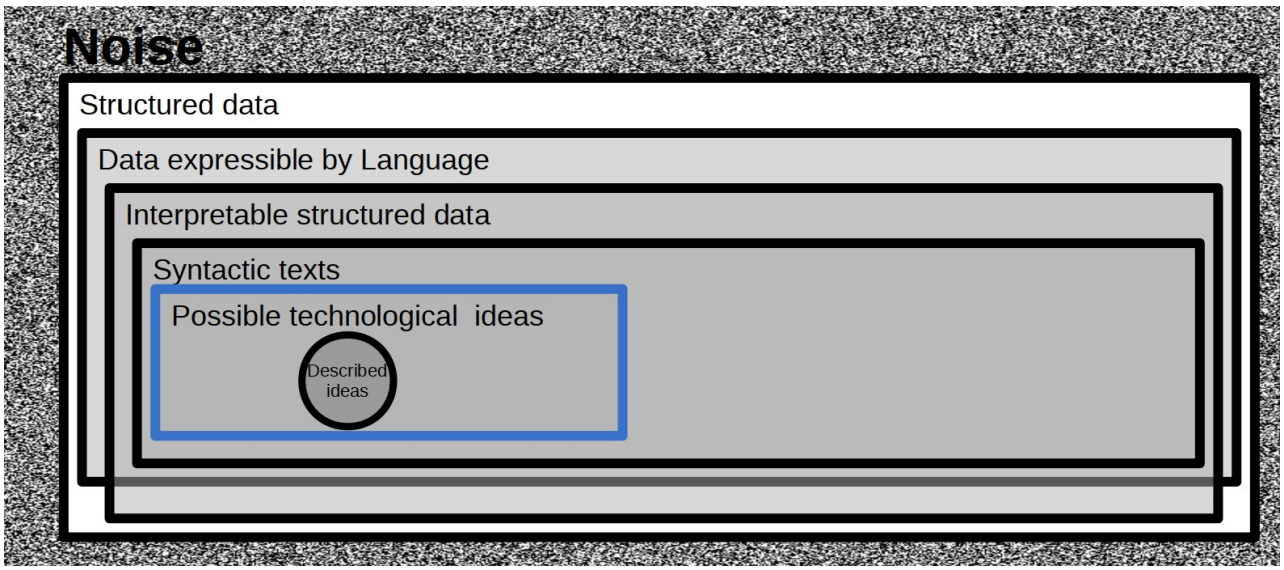

Illustration 9: Signal and noise.
Noise cannot be described by an algorithm, it cannot be compressed.

As for the question of automation of creativity, its "lower" rule-governed form seems to be a better candidate, although even that was inconceivable for Descartes: "We may easily conceive a machine to be so constructed that it emits vocables, […] but not that it should arrange them variously so as appositely to reply to what is said in its presence, as men of the lowest grade of intellect can do."[58] Others such as the Oulipo workshop of potential literature considered rules to be central to creativity.

One might consider that it is actually too easy to be creative without constraints. Constraints give structure and unstructured data is noise. Not mastering an instrument and creating a yet unheard noise with it is easy. Learning an instrument and playing a composition written by someone else may be difficult but is often considered to be less creative because the possible space of creation is very narrow, and even the goal of the performance is not necessarily novelty but rather pleasure or the expression of emotion. Creating a song that is new in the sense that it has never existed before while remaining recognizable as a song and pleasing to the ear is the form of creativity that might be the hardest to achieve. Yet, this is what Cloem does in the patent domain: The algorithm creates claim texts most of which have never existed before while remaining interpretable and recognizable as claim texts. This task is complex and resource-hungry from a computational point of view. It is not clear if rule-changing creativity, the creation of a whole new paradigm, triggering a Kuhnian paradigm shift, is actually more complex from a computational point of view. It is, though, certainly more complex from a social point of view – and humans do not seem ready to have their paradigms rewritten by an algorithm.

Just like for intelligence, we do not have satisfactory tests of when a machine has achieved creativity. The Turing Test remains a good attempt in this direction. So the equivalent for machine generated creative content is to ask whether the product can be distinguished from man-made creative content. The difficulty has always been not so much the creation of human-like creative

58 Original text: "on peut bien concevoir qu'une machine soit tellement faite qu'elle profère des paroles [...] mais non pas qu'elle les arrange diversement, pour répondre au sens de tout ce qui se dira en sa présence, ainsi que les hommes les plus hébétés peuvent faire." (Discourse on the Method, 1637)

content itself but rather the evaluation function and a feedback loop that strives to make the production indistinguishable from the original. A lot of research has been done on that question and recent machine learning techniques advance fast. Yet, for Cloem's shotgun approach, brutalizing the permutation space, the perfection of being indistinguishable is not actually needed as long as interesting and potentially relevant content is generated. Man-made patent texts are riddled with syntactic errors as many of them are written by non-natives, which does not deprive the texts of their juridical value.

An automatic machine learned estimation of the predicted quality of a patent text could be of use for ranking the search results in the Cloem database – a simple form of this is already implemented on the Cloem web page. If legal changes trying to protect the current patent system, that pose an upper limit to the number of ideas published together on the same page, were put into place, the ranking algorithm could provide a way of cutting off the less promising texts. But from a theoretical or implementation point of view, this is not necessary.

That this is possible can be seen as part of a process of demystification, the shrinking of human self-perception. It is a story of disenchantment. Just like the space of technological ideas is finite and the undiscovered space is shrinking, the space of novels, songs, and paintings is finite and the number of undiscovered objects is shrinking. And, more importantly, increasingly, machines explore systematically that unexplored world of ideas, selecting the best ones for humans to look at and enjoy.

What will the consequences be? We can observe what happened to human faculties other than creativity that were thought to be specifically human and that have ultimately been reached and completely taken over by machines, such as arithmetic or the capacity to accurately reproduce a scene. The result is that mental arithmetic has become a rather useless and deprecated skill that has its place more in freak shows than in daily life. Similarly, figurative art may remain a weird hobby, actual artists have invented impressionism and have long moved on from there to conceptual art. And even there, algorithms are moving in. The Swedish conceptual artist Jonas Lund[59] created an installation where a machine creates the titles of a new piece of art, including the materials to be used. The artist's role is then reduced to executing the machine's instructions.

As Summers-Stay (2011) explains, every time machines master a human skill, that skill loses its value in the human society. What will a society look like that esteems neither mental nor creative capacities? Science will surely fall from grace. Possible candidates for upcoming appreciation are morals – or the perfection of the human body itself.

59 http://postdigital.ens.fr/archives/portfolio/27112017

7 On Predictions, especially about the future[60]

“The past does not foretell the future” is a dictum that is often heard and yet goes against all scientific evidence, even against the very idea of science. Indeed, any theorem in physics as well as in psychology is acquired from the past and its validity is based on future events. The Popperian idea of falsification is a test of the predictive capacity of a theory.

At the same time, Chaos Theory shows us that there are systems of such complexity that no shortcut is possible: The best model for making a weather prediction is the very world that “calculates” tomorrow’s weather. Any simpler model would be imprecise. Yet, at least for a limited time span, predictions, aka weather forecasts, are possible in this system.

If we take the meaning of “foretelling” in the dictum as a precise and unique prediction, then, yes, even without the need to talk about chaos, the very inaccuracy of the measurements of the past prevents a perfect prediction, which thus becomes an oxymoron. If, on the other hand, a more vague prediction is allowed, into which fit a whole series of events, each validating the augury, as was the case with the Delphi oracle, it becomes much easier to predict – going as far as tautologies, exemplified by the rhyming German bon mot “*Wenn der Hahn kräht auf dem Mist, ändert sich das Wetter oder es bleibt wie’s ist”* ‘If the cock crows on the manure the weather will change or remain as it is now’.

Massive Defensive Publishing has to aim for the prediction of the future. A generated text is valuable if and only if its semantic space collides with a patent application. What makes this generation task so special is that we do not have to look for “the one” text, the system can process billions of variants. Yet, there are numerous reasons why we cannot rely on anything like typing monkeys, the first of which is the limitations of Cloem’s generation, storage, and query system that cannot handle such a notably higher number of variations. But even if the system could handle the sheer mass of texts, the system would not be usable for patent litigation, because the normal scenario of making use of Massive Defensive Publishing is as follows: A patent filing (application or grant) is identified as inconvenient. We search for existing published texts that collide with the inconvenient patent. For this, we look for expected keywords from the inconvenient patent text or their near-synonyms.

If we had typing monkeys instead of a smart algorithm for Massive Defensive Publishing, the database would overwhelmingly consist of meaningless letter combinations. A query with our keywords would, of course, return texts with these keywords, but the rest of the text would be meaningless gibberish, in nearly all texts from our database that contain the keywords. The other, interesting texts are there, too, but with an arbitrary ranking of the results, the chance of having them appear high up in the list are very low. Of course, we could imagine a ranking algorithm that puts grammatical sentences built from promising domain-specific vocabulary up front. But if we had a system that can do just that, why not use it before storage?

This is where Cloem and typing monkeys are completely different. Cloem is not about chance,

60 The title refers to the famous quote falsely attributed to Niels Bohr (https://quoteinvestigator.com/2013/10/20/no-predict/): *It is difficult to make predictions, especially about the future.*

about laws of large numbers. It is about optimizing grammar and vocabulary for a targeted semantic covering of semantic space where future patents are expected to appear. It resembles a beam-search algorithm in parsing technology where a balance has to be found between the available memory and the danger of cutting a branch off the search tree that contains the desired analysis. We have to implement when a text is a good prediction, considered as possibly interesting and non-tautological. In other words, we need to formalize how much of the possible, of the semantic space, can be affirmed, how much must be excluded for the augur to make sense, because too general vocabulary replacements (such as replacing each noun by *thing*) would result in tautological sentences in the sense that they collide with the inconveniently occupied semantic space, but are too general to be legally usable.

So one point we have to get right is the grammaticality of the sentences, we have to be able to parse the base claim, and we have to implement a grammar describing the allowed variations. The other point to tackle is the vocabulary. This is an even harder part because of the sheer magnitude of the constantly changing vocabulary of the technical domains.

Here an important insight from the weather forecasts comes in handy. The domain of inventions is probably just as complex as the weather but just like for the weather, we can make good short-term predictions. These predictions are never perfect in the chaotic system, but we can estimate the likeliness of predictions, a probability that is declining with the distance of the prediction.

Exceptions exist but generally each vocable follows typical trajectories over its lifetime, from its creation as a new term, to fame, to oblivion. These patterns are detectable in patent and other technological texts. This is the bet that we made when developing the Hackathon of the TALN 2017 conference, called HackaTAL[61], together with Damien Nouvel. The participants were asked to employ machine learning techniques on a very large database of patent texts, organized by domains, in order to detect emerging themes.

The results, code and data on the Hackathon's website, look promising, but are difficult to evaluate. One path that remains to be tested systematically is moving back in time: For a given technological domain in a given year, for instance A23L3 (Methods for heating or cooling material in a container) in the year 2011, we identify a set of specific terms, words that are "trendy" that year and should thus have been used had we done Massive Defensive Publishing in the year 2010. Take *mayonnaise* to be one of the words to find. We now have a classical machine learning problem: How to predict *mayonnaise* from the trend lines of all the terms from the, say, 10 years of patents in the A23L3 field prior to 2011, 2000-2010? And more generally, for each year and each domain we want to predict all the specific terms from the years preceding their specificity. In a word, we want to learn to detect weak signals in the short-term diachronic development of word usage. This could significantly improve the predictive force of the Cloem variants of a base claim. For the moment the Cloem variation vocabulary is based on simpler measures of general synchronic word usage and combinatorics.

Since this approach has already given promising results on simple vocabulary, there is no reason that it should not work on the prediction of more complex expressions and grammatical sentence

61 https://github.com/HackaTAL/2017

structures.

The effectiveness of such an approach to vocabulary and sentence structure selection for Massive Defensive Publishing is based on an uncanny underlying assumption: Even inventions are not completely surprising but, to some extent, predictable. Inventors are not completely free but their actual choices are limited. And this limitation concerns not only the expected stable grammar but also the vocabulary and expressions they use. Even inventors are better if they follow the trends, if they are not too inventive, because their ideas may land too far off existing technologies, resembling unstructured irrelevant noise.

This brings us back to the theme that has accompanied us all along this text: Technology keeps reminding us that the dreamed up infinity of human language boils down to a handful of choices, that Borges' Library of Babel is just a few shelves long, and what appeared to be language creativity shrinks to strictly following rules, to staying on the well beaten path. And linguistics, any science in fact, just like Language itself, loves repetition; unique events are noise, noise a scientist shamefacedly turns away from, because seeing unique events means that we have not been able to detect the hidden regularity.

Yet, Russell's Inductivist Turkey, the turkey that believed in the linear regression analysis of his weight gain until Christmas Day, teaches us that trend detection will work, except when it won't, when the revolution comes, the paradigm changes.

If, however, we dare to prolong a few trend lines in linguistics, we can predict certain things, alas with an unknown probability:

We will foresee that measures and measurement tools with predictive algorithms will reach a new level of omnipresence in our lives, waking us up when we have slept enough and telling us what to eat and to read, whom to love and to vote for, when to go to bed and to give up hope. Measures and machine predictions will substitute direct perceptions where the sensations deceive us.

Like microscopes in biology, machine learning and textual statistics tools will become part of the daily routing of the linguist. The linguist will get used to some tools up to the point of no longer realizing that the observation is mediated.

On a longer term, humans will possibly lose the need to do science because all ideas that our cognition can hold will soon have been explored. Yet, for the time being, I would suspect that the fascination for our language faculty will persist, even if a finite list of sentence is all a language is, because the optimal description of the finite list is rule-governed, at least for parts of the grammar.

We will predict that even formal grammars will survive the assault of machine learning, but only as scientific or pedagogical toys, downgraded from empirical engineering tools optimized for prediction. This will allow us to become aware that all science is turning into Human Sciences, because science is something we do for pleasure and for pleasure only.

8 My publications

9 book editions, 6 chapters in books, 6 journal articles, 50 articles in peer-reviewed conference proceedings.

8.1 Organized chronologically in descending order

1. **Book:** Kahane, Sylvain, and Gerdes, Kim. *Syntaxe*. Language Science Press. In preparation.
2. **Journal Article:** Gerdes, Kim, Sylvain Kahane, and Xinying Chen. Typometrics From Implicational to Quantitative Universals in Word Order Typology, submitted.
3. **Chapter:** Kahane, Sylvain, Kim Gerdes, and Rachel Bawden. "The microsyntactic annotation", *A Prosodic and Syntactic Treebank for Spoken French*, John Benjamins, Amsterdam. In press.
4. **Chapter:** Kahane, Sylvain, Paola Pietrandrea, and Kim Gerdes. "The annotation of list structures". *A Prosodic and Syntactic Treebank for Spoken French*, John Benjamins, Amsterdam. In press.
5. **Chapter:** Gerdes, Kim, Sylvain Kahane, Julie Beliao, and Éric de la Clergerie. "Annotation tools for syntax". *A Prosodic and Syntactic Treebank for Spoken French*, John Benjamins, Amsterdam. In press.
6. **Journal Article:** Osborne, Tim, Kim Gerdes. The Status of Function Words in Dependency Grammar: A Critique of Universal Dependencies (UD), to appear in *Glossa*. 2018.
7. **Proceedings Article:** Gerdes, Kim, Bruno Guillaume, Sylvain Kahane, Guy Perrier. "SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD". *Proceedings of the Universal Dependencies Workshop 2018*.
8. **Chapter:** Chen, Xinying, and Kim Gerdes. "How Do Universal Dependencies Distinguish Language Groups?" *Quantitative Analysis of Dependency Structures*, Quantitative Linguistics [QL] 72, Pages 277-294. De Gruyter Mouton. 2018.
9. **Proceedings Article:** Courtin, Marine, Bernard Caron, Kim Gerdes, and Sylvain Kahane. "Establishing a Language by Annotating a Corpus: the Case of Naija, a Post-creole Spoken in Nigeria" Proceedings of the Workshop on Annotation in Digital Humanities, 2018.
10. **Video:** Gerdes, Kim. "L'exploration mécanique du possible : de la divination à l'écriture de toutes les idées" *Public debate with Michel de Broin in the Postdigital seminar on Machines organiques, machines textuelles.* École normale supérieure, Paris. 26/03/2018. https://www.youtube.com/watch?v=5I9YMwxOY6k
11. **Proceedings Article:** Kahane, Sylvain, Marine Courtin, and Kim Gerdes. "Multi-word annotation in syntactic treebanks-Propositions for Universal Dependencies" In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pp. 181-189. 2018.
12. **Proceedings Article:** Chen, Xinying, and Kim Gerdes. "Classifying Languages by Dependency Structure. Typologies of Delexicalized Universal Dependency Treebanks" In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), September 18-20, 2017, Università di Pisa, Italy*, no. 139, pp. 54-63. Linköping University Electronic Press, 2017.

13. **Proceedings Article:** Kahane, Sylvain, Henri-José Deulofeu, Kim Gerdes, Alexis Nasr, and André Valli. "Annotation micro-et macrosyntaxique manuelle et automatique de français parlé" In Proceedings of the *Journée Floral*. 2017.
14. **Proceedings Article:** Gerdes, Kim, and Sylvain Kahane. "Trois schémas d'annotation syntaxique en dépendance pour un même corpus de français oral: le cas de la macrosyntaxe" In *Atelier sur les corpus annotés du français (ACor4French)*. 2017.
15. **Proceedings Article:** Wong, Tak-sum, Kim Gerdes, Herman Leung, and John Lee. "Quantitative Comparative Syntax on the Cantonese-Mandarin Parallel Dependency Treebank" In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pp. 266-275. 2017.
16. **Proceedings Article:** Leung, Herman, Rafaël Poiret, Tak-sum Wong, Xinying Chen, Kim Gerdes, and John Lee. "Developing Universal Dependencies for Mandarin Chinese" In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pp. 20-29. 2016.
17. **Journal Article:** Gendrot, Cédric, Kim Gerdes, and Martine Adda-Decker. "Détection automatique d'une hiérarchie prosodique dans un corpus de parole journalistique." *Langue française* 3 (2016): 123-149.
18. **Technical Report:** Kahane, Sylvain, Henri-José Deulofeu, Kim Gerdes, André Valli, and Alexis Nasr. "Guide d'annotation syntaxique Orféo (version Platinum)". 2016.
19. **Proceedings Article:** Gerdes, Kim, and Sylvain Kahane. "Dependency Annotation Choices: Assessing Theoretical and Practical Issues of Universal Dependencies" In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pp. 131-140. 2016.
20. **Proceedings Article:** Guibon, Gaël, Isabelle Tellier, Sophie Prévost, Mathieu Constant, and Kim Gerdes. "Searching for Discriminative Metadata of Heterogenous Corpora" In *Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*, pp. 72-82. 2015.
21. **Proceedings Article:** Guibon, Gael, Isabelle Tellier, Sophie Prévost, Matthieu Constant, and Kim Gerdes. "Analyse syntaxique de l'ancien français: quelles propriétés de la langue influent le plus sur la qualité de l'apprentissage?" In *Proceedings of TALN* 2015.
22. **Proceedings Article:** Gerdes, Kim, and Sylvain Kahane. "Non-constituent coordination and other coordinative constructions as Dependency Graphs" In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pp. 101-110. 2015.
23. **Proceedings Article:** Chen, Xinying, Haitao Liu, and Kim Gerdes. "Classifying Syntactic Categories in the Chinese Dependency Network" In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pp. 74-81. 2015.
24. **Proceedings Article:** Guibon, Gaël, Isabelle Tellier, Mathieu Constant, Sophie Prévost, and Kim Gerdes. "Parsing Poorly Standardized Language Dependency on Old French" In *Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)*, pp. 51-61. 2014.
25. **Edition:** Gerdes, Kim, Eva Hajičová, and Leo Wanner (eds). *Dependency Linguistics: Recent advances in linguistic theory using dependency structures*. Vol. 215. Studies in Language Companion Series. John Benjamins Publishing Company, 2014.

26. **Proceedings Article:** Lacheret, Anne, Sylvain Kahane, Julie Beliao, Anne Dister, Kim Gerdes, Jean-Philippe Goldman, Nicolas Obin, Paola Pietrandrea, and Atanas Tchobanov. "Rhapsodie: a Prosodic-Syntactic Treebank for Spoken French" In *Language Resources and Evaluation Conference*. 2014.
27. **Proceedings Article:** Lacheret, Anne, Sylvain Kahane, Julie Beliao, Anne Dister, Kim Gerdes, Jean-Philippe Goldman, Nicolas Obin, Paola Pietrandrea, and Atanas Tchobanov. "Rhapsodie: un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé" In Proceedings of the *4th World Congress of French Linguistics (CMLF) - SHS Web of Conferences*, vol. 8, pp. 2675-2689. EDP Sciences, 2014.
28. **Proceedings Article:** Bawden, Rachel, Marie-Amélie Bottala, Kim Gerdes, and Sylvain Kahane. "Correcting and Validating Syntactic Dependency in the Spoken French Treebank Rhapsodie" In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, pp. 1-6. 2014.
29. **Proceedings Article:** Gerdes, Kim. "Corpus collection and analysis for the linguistic layman: The Gromoteur" In *Proceedings of JADT*. 2014.
30. **Edition:** Gerdes, Kim, Eva Hajičová, and Leo Wanner, eds. *Computational Dependency Theory*. Vol. 258. IOS Press, 2013.
31. **Proceedings Article:** Gerdes, Kim. "Collaborative dependency annotation" *Proceedings of the second international conference on dependency linguistics (DepLing 2013)*. 2013.
32. **Edition:** Hajičová, Eva, Kim Gerdes, and Leo Wanner. *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*. 2013.
33. **Proceedings Article:** Gerdes, Kim, Sylvain Kahane, Anne Lacheret, Arthur Truong, and Paola Pietrandrea. "Intonosyntactic data structures: the Rhapsodie treebank of spoken French" In *Proceedings of the Sixth Linguistic Annotation Workshop*, pp. 85-94. Association for Computational Linguistics, 2012.
34. **Proceedings Article:** Gendrot, Cédric, Kim Gerdes, and Martine Adda-Decker. "Impact of prosodic position on vocalic space in German and French" In *Proceedings of the 17th international congress of phonetic sciences*, pp. 731-734. 2011.
35. **Proceedings Article:** Gerdes, Kim, and Sylvain Kahane. 2011. "Defining dependencies (and constituents)", Proceedings of Depling 2011. Long version in *Frontiers in Artificial Intelligence and Applications* 258 (2013): 1-25.
36. **Edition:** Gerdes, Kim, Eva Hajičová, and Leo Wanner. *Proceedings of the Dependency Linguistics 2011 conference (Depling 2011)*. 2011. Introduction.
37. **Chapter:** Miao, Jun, and Kim Gerdes. "进入傅雷译作" 'Giving access to Fu Lei's work' 傅雷的精神世界及其时代意义 *'Fu Lei's Spiritual World in its significance'*, 中西书局 'Chinese Western Book Edition', Shanghai, pp. 351-366. 2011.
38. **Proceedings Article:** Deulofeu, José, Lucie Duffort, Kim Gerdes, Sylvain Kahane, and Paola Pietrandrea. "Depends on what the French say: spoken corpus annotation with and beyond syntactic functions" In *Proceedings of the Fourth Linguistic Annotation Workshop*, pp. 274-281. Association for Computational Linguistics, 2010.
39. **Proceedings Article:** Benzitoun, C., Dister, A., Gerdes, K., Kahane, S., Pietrandrea, P. and Sabio, F., 2010, July. tu veux couper là faut dire pourquoi-Propositions pour une segmentation syntaxique du français parlé. In *Congrès Mondial de linguistique française* (pp. 2075-2090). CMLF.

40. **Proceedings Article:** Gendrot, Cédric, and Kim Gerdes. "Hiérarchie prosodique et réalisation spectrale des voyelles" In *XXVIIIèmes Journées d'Étude sur la Parole*, pp. 85-88. 2010.
41. **Proceedings Article:** Gendrot, Cédric, and Kim Gerdes. "Prosodic boundaries and spectral realization of French vowels" In *GLOW Workshop on Phonology and Phonetics*. 2010.
42. **Proceedings Article:** Benzitoun, Christophe, Anne Dister, Kim Gerdes, Sylvain Kahane, and Renaud Marlet. "annoter du des textes tu te demandes si c'est syntaxique tu vois." In *28th International Conference on Lexis and Grammar (LGC 2009)*, vol. 4, pp. 16-27. Presses de l'Université de Bergen, 2009.
43. **Proceedings Article:** Gendrot, Cédric, and Kim Gerdes. "Prosodic Hierarchy and Spectral Realization of Vowels in French" *Interface Discours & Prosodie*: 191-205. 2009.
44. **Proceedings Article:** Gerdes, Kim, Renaud Marlet, and Lionel Clément. "A Grammar Correction Algorithm: Deep Parsing and Minimal Corrections for a Grammar Checker" In *Formal Grammar. 14th International Conference, FG 2009*.
45. **Proceedings Article:** Clément, Lionel, Kim Gerdes, and Renaud Marlet. "Grammaires d'erreur–correction grammaticale avec analyse profonde et proposition de corrections minimales" In *16e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2009)*, p. electronique. 2009.
46. **Proceedings Article:** Magistry, Pierre, and Kim Gerdes. "Paddy Fields: A Topological Description of Chinese Word Order" *Proceedings Fourth International Conference on Meaning-Text Theory*. 2009.
47. **Edition:** Gerdes, Kim, David Beck, Jasmina Milićević, and Alain Polguère. *Proceedings of the 4th International Conference on Meaning-Text Theory*: 410 pages. 2009.
48. **Proceedings Article:** Gerdes, Kim, and Sylvain Kahane. "Speaking in Piles: Paradigmatic Annotation of a French Spoken Corpus" *Proceedings of the Fifth Corpus Linguistics Conference, Liverpool*. 2009.
49. **Proceedings Article:** Gerdes, Kim. "L'alignement pour les pauvres: Adapter la bonne métrique pour un algorithme dynamique de dilatation temporelle pour l'alignement sans ressources de corpus bilingues" *Proceedings of the 9es Journées internationales d'Analyse statistique des Données Textuelles* (2008a): 527-538.
50. **Proceedings Article:** Gerdes, Kim. "Poverty driven bilingual alignment" In *Proceedings of the International Symposium of Using Corpora in Contrastive and Translation Studies*. 2008b.
51. **Proceedings Article:** Gerdes, Kim, and Pollet Samvelian. "A statistical approach to Persian light verb constructions" Proceedings of the 27th Conference on Lexis and Grammar. 2008.
52. **Chapter:** Gerdes, Kim, and Sylvain Kahane. "Phrasing it differently" in: *Selected Lexical and Grammatical Issues in the Meaning Text Theory*. 2007.
53. **Edition:** Gerdes, Kim, Tilmann Reuther, and Leo Wanner (eds.). *MTT 2007: Meaning-Text Theory 2007: Proceedings of the 3rd International Conference on Meaning-Text Theory*, Klagenfurt, May 20-24, 2007.

54. **Proceedings Article:** Gerdes, Kim, and Sylvain Kahane. "A Polynomial Parsing Algorithm for the Topological Model: Synchronizing Constituent and Dependency Grammars, Illustrated by German Word Order Phenomena" In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 1097-1104. Association for Computational Linguistics, 2006.
55. **Proceedings and Journal Article:** Gerdes, Kim, and Sylvain Kahane. "L'amas verbal au cœur d'une modélisation topologique du français" *Proceedings of the Journées de la syntaxe–Ordre des mots dans la phrase française, positions et topologie* (2004), long version in *Lingvisticae Investigationes*, *29*(1), 75-89. 2006.
56. **Proceedings Article:** Clément, Lionel, and Kim Gerdes. "Analyzing zeugmas in XLFG" In *Proceedings of LFG 2006*. 2006.
57. **Edition:** Gerdes, Kim, and Claude Muller. *Actes des 4èmes Journées de Syntaxe de l'ERSS : Ordre des mots dans la phrase française, positions et topologie*. *Linguisticae Investigationes*, 29-1. Introduction. 2006.
58. **Journal Article:** Gerdes, Kim. "Sur la non-équivalence des représentations syntaxiques: Comment la représentation en X-barre nous amène au concept du mouvement" *Cahiers de grammaire* 30: 175-192. 2006.
59. **Proceedings Article:** Gerdes, Kim. "German Partial VP-fronting in a Meaning-Text Approach" Workshop Sentence-initial and sentence-final positions on the interplay between syntax, semantics and information structure. Abstract in the *Proceedings of the 38th meeting of the Societas Linguistica Europaea,* Valencia. 2005.
60. **Proceedings Article:** Gerdes, Kim, Sylvain Kahane, and Hi-Yon Yoo. "On the descriptive adequacy of topology" *Proceedings of MTT'05*. 2005.
61. **Journal Article:** Gerdes, Kim. "Tree Unification Grammar: Problems and Proposals for Topology, TAG, and German" *Electronic Notes in Theoretical Computer Science* 53: 108-136. 2004.
62. **Proceedings Article:** Gerdes, Kim, and Hi-Yon Yoo. "The fields on the way to prosody: Alternatives to phrase structure based approaches to prosody" In *Proceedings of the International Conference on Phonetic Sciences*. 2003.
63. **Proceedings Article:** Gerdes, Kim, and Hi-Yon Yoo. "La topologie comme interface entre syntaxe et prosodie: un système de génération appliqué au grec moderne." *Proceedings of TALN 2003* (2003): 125-134.
64. **Proceedings Article:** Gerdes, Kim, and Hi-Yon Yoo. "The fields on the way to prosody Alternatives to phrase structure based approaches to prosody." *Proceedings of the International Conference on Phonetic Sciences*. 2003.
65. **Proceedings Article:** Gerdes, Kim, and Hi-Yon Yoo. "A dependency account of Korean word order", *Proceedings of the International Conference of the Linguistic Society of Korea*, Seoul. 2003.
66. **Proceedings Article:** Gerdes, Kim, and Hi-Yon Yoo. "Greek Topology: A Prosodic Generation System for Modern Greek". *Proceedings of the Workshop on Prosodic Interfaces / Interfaces Prosodiques*, Nantes, France. 2003
67. **Proceedings Article:** Gerdes, Kim. "DTAG?" *Proceedings of the sixth international workshop on tree adjoining grammar and related frameworks (TAG+6)*. 2002.

68. **Thesis:** Gerdes, K. *Topologie et grammaires formelles de l'allemand* (Doctoral dissertation, Paris 7). 2002.

69. **Proceedings Article:** Clément, Lionel, Kim Gerdes, and Sylvain Kahane. "An LFG-type grammar for German based on the topological model" *Proceedings of LFG 2002*.

70. **Proceedings Article:** Gerdes, Kim, and Sylvain Kahane. "Word order in German: A formal dependency grammar using a topological hierarchy" In *Proceedings of the 39th annual meeting on association for computational linguistics*, pp. 220-227. Association for Computational Linguistics, 2001.

71. **Proceedings Article:** Gerdes Kim & Kahane Sylvain. "A Description of German Syntax Based on a Topological Hierarchy", *Proceedings of the 7th Germanic Linguistics Annual Conference (GLAC'7)*. 2001.

72. **Proceedings Article:** Gerdes, Kim, and Sylvain Kahane 2000. "Pas de syntaxe sans prosodie : illustration par l'allemand", *Proceedings of the XXIIIèmes Journées d'Etude sur la Parole*, Aussois, 19-23 juin 2000.

73. **Proceedings Article:** Gerdes, Kim, and Patrice Lopez. "Shared Semantic Dependency Representations for LTAG" *Proceedings of the Journée ATALA Représentation et traitement de l'ambiguité pour l'analyse syntaxique*. 2000.

74. **Thesis:** Gerdes, Kim. *Le cas allemand en TAG* Master Thesis, University Paris 7. 1998.

75. **Thesis:** Gerdes, Kim and Julia Kempe. *Décomposition équidimensionnelle de variétés en temps polynomial* Master Thesis. École polytechnique and University Paris XI. 1996.

8.2 Organized by Themes

8.2.1 On formal grammar and topology

Proceedings Article: Magistry, Pierre, and Kim Gerdes. "Paddy Fields: A Topological Description of Chinese Word Order" *Proceedings Fourth International Conference on Meaning-Text Theory*. 2009.

Chapter: Gerdes, Kim, and Sylvain Kahane. "Phrasing it differently" in: *Selected Lexical and Grammatical Issues in the Meaning Text Theory*. 2007.

Proceedings Article: Gerdes, Kim, and Sylvain Kahane. "A Polynomial Parsing Algorithm for the Topological Model: Synchronizing Constituent and Dependency Grammars, Illustrated by German Word Order Phenomena" In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 1097-1104. Association for Computational Linguistics, 2006.

Proceedings and Journal Article: Gerdes, Kim, and Sylvain Kahane. "L'amas verbal au cœur d'une modélisation topologique du français" *Proceedings of the Journées de la syntaxe–Ordre des mots dans la phrase française, positions et topologie* (2004), long version in *Lingvisticae Investigationes*, *29*(1), 75-89. 2006.

Proceedings Article: Gerdes, Kim, Sylvain Kahane, and Hi-Yon Yoo. "On the descriptive adequacy of topology" *Proceedings of MTT'05*. 2005.

Journal Article: Gerdes, Kim. "Tree Unification Grammar: Problems and Proposals for Topology, TAG, and German" *Electronic Notes in Theoretical Computer Science* 53: 108-136. 2004.

Proceedings Article: Gerdes, Kim, and Hi-Yon Yoo. "The fields on the way to prosody: Alternatives to phrase structure based approaches to prosody" In *Proceedings of the International Conference on Phonetic Sciences*. 2003.

Proceedings Article: Gerdes, Kim, and Hi-Yon Yoo. "La topologie comme interface entre syntaxe et prosodie: un système de génération appliqué au grec moderne." *Proceedings of TALN 2003* (2003): 125-134.

Proceedings Article: Gerdes, Kim, and Hi-Yon Yoo. "The fields on the way to prosody Alternatives to phrase structure based approaches to prosody." *Proceedings of the International Conference on Phonetic Sciences*. 2003.

Proceedings Article: Gerdes, Kim, and Hi-Yon Yoo. "A dependency account of Korean word order", *Proceedings of the International Conference of the Linguistic Society of Korea*, Seoul. 2003.

Proceedings Article: Gerdes, Kim, and Hi-Yon Yoo. "Greek Topology: A Prosodic Generation System for Modern Greek". *Proceedings of the Workshop on Prosodic Interfaces / Interfaces Prosodiques*, Nantes, France. 2003

Proceedings Article: Gerdes, Kim. "DTAG?" *Proceedings of the sixth international workshop on tree adjoining grammar and related frameworks (TAG+6)*. 2002.

Thesis: Gerdes, K. *Topologie et grammaires formelles de l'allemand* (Doctoral dissertation, Paris 7). 2002.

Proceedings Article: Clément, Lionel, Kim Gerdes, and Sylvain Kahane. "An LFG-type grammar for German based on the topological model" *Proceedings of LFG 2002*.

Proceedings Article: Gerdes, Kim, and Sylvain Kahane. "Word order in German: A formal dependency grammar using a topological hierarchy" In *Proceedings of the 39th annual meeting on association for computational linguistics*, pp. 220-227. Association for Computational Linguistics, 2001.

Proceedings Article: Gerdes Kim & Kahane Sylvain. "A Description of German Syntax Based on a Topological Hierarchy", *Proceedings of the 7th Germanic Linguistics Annual Conference (GLAC'7)*. 2001.

Proceedings Article: Gerdes, Kim, and Sylvain Kahane 2000. "Pas de syntaxe sans prosodie : illustration par l'allemand", *Proceedings of the XXIIIèmes Journées d'Etude sur la Parole*, Aussois, 19-23 juin 2000.

Thesis: Gerdes, Kim. Le cas allemand en TAG Master Thesis, University Paris 7. 1998.

8.2.2 On other syntactic descriptions

Book: Kahane, Sylvain, and Gerdes, Kim. *Syntaxe*. Language Science Press. In preparation.

Proceedings Article: Gerdes, Kim, and Sylvain Kahane. "Non-constituent coordination and other coordinative constructions as Dependency Graphs" In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pp. 101-110. 2015.

Proceedings Article: Gerdes, Kim, and Sylvain Kahane. 2011. "Defining dependencies (and constituents)", Proceedings of Depling 2011. Long version in *Frontiers in Artificial Intelligence and Applications* 258 (2013): 1-25.

Proceedings Article: Clément, Lionel, and Kim Gerdes. "Analyzing zeugmas in XLFG" In *Proceedings of LFG 2006*. 2006.

Proceedings Article: Gerdes, Kim. "German Partial VP-fronting in a Meaning-Text Approach" Workshop Sentence-initial and sentence-final positions on the interplay between syntax, semantics and information structure. Abstract in the *Proceedings of the 38th meeting of the Societas Linguistica Europaea*, Valencia. 2005.

Proceedings Article: Gerdes, Kim, and Patrice Lopez. "Shared Semantic Dependency Representations for LTAG" *Proceedings of the Journée ATALA Représentation et traitement de l'ambiguité pour l'analyse syntaxique*. 2000.

8.2.3 On corpus annotation

Chapter: Kahane, Sylvain, Kim Gerdes, and Rachel Bawden. "The microsyntactic annotation", *A Prosodic and Syntactic Treebank for Spoken French*, John Benjamins, Amsterdam. In press.

Chapter: Kahane, Sylvain, Paola Pietrandrea, and Kim Gerdes. "The annotation of list structures". *A Prosodic and Syntactic Treebank for Spoken French*, John Benjamins, Amsterdam. In press.

Journal Article: Osborne, Tim, Kim Gerdes. The Status of Function Words in Dependency Grammar: A Critique of Universal Dependencies (UD), to appear in *Glossa*. 2018.

Proceedings Article: Gerdes, Kim, Bruno Guillaume, Sylvain Kahane, Guy Perrier. "SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD". *Proceedings of the Universal Dependencies Workshop 2018*.

Proceedings Article: Courtin, Marine, Bernard Caron, Kim Gerdes, and Sylvain Kahane. "Establishing a Language by Annotating a Corpus: the Case of Naija, a Post-creole Spoken in Nigeria" Proceedings of the Workshop on Annotation in Digital Humanities, 2018.

Proceedings Article: Kahane, Sylvain, Marine Courtin, and Kim Gerdes. "Multi-word annotation in syntactic treebanks-Propositions for Universal Dependencies" In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pp. 181-189. 2018.

Technical Report: Kahane, Sylvain, Henri-José Deulofeu, Kim Gerdes, André Valli, and Alexis Nasr. "Guide d'annotation syntaxique Orféo (version Platinum)". 2016.

Proceedings Article: Kahane, Sylvain, Henri-José Deulofeu, Kim Gerdes, Alexis Nasr, and André Valli. "Annotation micro-et macrosyntaxique manuelle et automatique de français parlé" In Proceedings of the *Journée Floral*. 2017.

Proceedings Article: Gerdes, Kim, and Sylvain Kahane. "Trois schémas d'annotation syntaxique en dépendance pour un même corpus de français oral: le cas de la macrosyntaxe" In *Atelier sur les corpus annotés du français (ACor4French)*. 2017.

Proceedings Article: Leung, Herman, Rafaël Poiret, Tak-sum Wong, Xinying Chen, Kim Gerdes, and John Lee. "Developing Universal Dependencies for Mandarin Chinese" In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pp. 20-29. 2016.

Proceedings Article: Gerdes, Kim, and Sylvain Kahane. "Dependency Annotation Choices: Assessing Theoretical and Practical Issues of Universal Dependencies" In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pp. 131-140. 2016.

Proceedings Article: Guibon, Gaël, Isabelle Tellier, Mathieu Constant, Sophie Prévost, and Kim Gerdes. "Parsing Poorly Standardized Language Dependency on Old French" In *Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)*, pp. 51-61. 2014.

Proceedings Article: Lacheret, Anne, Sylvain Kahane, Julie Beliao, Anne Dister, Kim Gerdes, Jean-Philippe Goldman, Nicolas Obin, Paola Pietrandrea, and Atanas Tchobanov. "Rhapsodie: a Prosodic-Syntactic Treebank for Spoken French" In *Language Resources and Evaluation Conference*. 2014.

Proceedings Article: Lacheret, Anne, Sylvain Kahane, Julie Beliao, Anne Dister, Kim Gerdes, Jean-Philippe Goldman, Nicolas Obin, Paola Pietrandrea, and Atanas Tchobanov. "Rhapsodie: un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé" In Proceedings of the *4th World Congress of French Linguistics (CMLF) - SHS Web of Conferences*, vol. 8, pp. 2675-2689. EDP Sciences, 2014.

Proceedings Article: Bawden, Rachel, Marie-Amélie Bottala, Kim Gerdes, and Sylvain Kahane. "Correcting and Validating Syntactic Dependency in the Spoken French Treebank Rhapsodie" In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, pp. 1-6. 2014.

Proceedings Article: Gerdes, Kim, Sylvain Kahane, Anne Lacheret, Arthur Truong, and Paola Pietrandrea. "Intonosyntactic data structures: the Rhapsodie treebank of spoken French" In *Proceedings of the Sixth Linguistic Annotation Workshop*, pp. 85-94. Association for Computational Linguistics, 2012.

Proceedings Article: Deulofeu, José, Lucie Duffort, Kim Gerdes, Sylvain Kahane, and Paola Pietrandrea. "Depends on what the French say: spoken corpus annotation with and beyond syntactic functions" In *Proceedings of the Fourth Linguistic Annotation Workshop*, pp. 274-281. Association for Computational Linguistics, 2010.

Proceedings Article: Benzitoun, C., Dister, A., Gerdes, K., Kahane, S., Pietrandrea, P. and Sabio, F., 2010, July. tu veux couper là faut dire pourquoi-Propositions pour une segmentation syntaxique du français parlé. In *Congrès Mondial de linguistique française* (pp. 2075-2090). CMLF.

Proceedings Article: Benzitoun, Christophe, Anne Dister, Kim Gerdes, Sylvain Kahane, and Renaud Marlet. "annoter du des textes tu te demandes si c'est syntaxique tu vois." In *28th International Conference on Lexis and Grammar (LGC 2009)*, vol. 4, pp. 16-27. Presses de l'Université de Bergen, 2009.

Proceedings Article: Gerdes, Kim, and Sylvain Kahane. "Speaking in Piles: Paradigmatic Annotation of a French Spoken Corpus" *Proceedings of the Fifth Corpus Linguistics Conference, Liverpool*. 2009.

8.2.4 On corpus exploitation

Chapter: Gerdes, Kim, Sylvain Kahane, and Xinying Chen. Typometrics From Implicational to Quantitative Universals in Word Order Typology, submitted.

Chapter: Chen, Xinying, and Kim Gerdes. "How Do Universal Dependencies Distinguish Language Groups?" *Quantitative Analysis of Dependency Structures*, Quantitative Linguistics [QL] 72, Pages 277-294. De Gruyter Mouton. 2018.

Proceedings Article: Chen, Xinying, and Kim Gerdes. "Classifying Languages by Dependency Structure. Typologies of Delexicalized Universal Dependency Treebanks" In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), September 18-20, 2017, Università di Pisa, Italy*, no. 139, pp. 54-63. Linköping University Electronic Press, 2017.

Proceedings Article: Wong, Tak-sum, Kim Gerdes, Herman Leung, and John Lee. "Quantitative Comparative Syntax on the Cantonese-Mandarin Parallel Dependency Treebank" In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pp. 266-275. 2017.

Journal Article: Gendrot, Cédric, Kim Gerdes, and Martine Adda-Decker. "Détection automatique d'une hiérarchie prosodique dans un corpus de parole journalistique." *Langue française* 3 (2016): 123-149.

Proceedings Article: Guibon, Gaël, Isabelle Tellier, Sophie Prévost, Mathieu Constant, and Kim Gerdes. "Searching for Discriminative Metadata of Heterogenous Corpora" In *Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*, pp. 72-82. 2015.

Proceedings Article: Guibon, Gael, Isabelle Tellier, Sophie Prévost, Matthieu Constant, and Kim Gerdes. "Analyse syntaxique de l'ancien français: quelles propriétés de la langue influent le plus sur la qualité de l'apprentissage?" In *Proceedings of TALN* 2015.

Proceedings Article: Chen, Xinying, Haitao Liu, and Kim Gerdes. "Classifying Syntactic Categories in the Chinese Dependency Network" In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pp. 74-81. 2015.

Proceedings Article: Gendrot, Cédric, Kim Gerdes, and Martine Adda-Decker. "Impact of prosodic position on vocalic space in German and French" In *Proceedings of the 17th international congress of phonetic sciences*, pp. 731-734. 2011.

Chapter: Miao, Jun, and Kim Gerdes. "进入傅雷译作" 'Giving access to Fu Lei's work' 傅雷的精神世界及其时代意义 *'Fu Lei's Spiritual World in its significance'*, 中西书局 'Chinese Western Book Edition', Shanghai, pp. 351-366. 2011.

Proceedings Article: Gendrot, Cédric, and Kim Gerdes. "Hiérarchie prosodique et réalisation spectrale des voyelles" In *XXVIIIèmes Journées d'Étude sur la Parole*, pp. 85-88. 2010.

Proceedings Article: Gendrot, Cédric, and Kim Gerdes. "Prosodic boundaries and spectral realization of French vowels" In *GLOW Workshop on Phonology and Phonetics*. 2010.

Proceedings Article: Gendrot, Cédric, and Kim Gerdes. "Prosodic Hierarchy and Spectral Realization of Vowels in French" *Interface Discours & Prosodie*: 191-205. 2009.

Proceedings Article: Gerdes, Kim, and Pollet Samvelian. "A statistical approach to Persian light verb constructions" Proceedings of the 27th Conference on Lexis and Grammar. 2008.

8.2.5 On NLP tools (annotation, parsing, alignment, ...)

Chapter: Gerdes, Kim, Sylvain Kahane, Julie Beliao, and Éric de la Clergerie. "Annotation tools for syntax". *A Prosodic and Syntactic Treebank for Spoken French*, John Benjamins, Amsterdam. In press.

Proceedings Article: Gerdes, Kim. "Corpus collection and analysis for the linguistic layman: The Gromoteur" In *Proceedings of JADT*. 2014.

Proceedings Article: Gerdes, Kim. "Collaborative dependency annotation" *Proceedings of the second international conference on dependency linguistics (DepLing 2013)*. 2013.

Proceedings Article: Gerdes, Kim, Renaud Marlet, and Lionel Clément. "A Grammar Correction Algorithm: Deep Parsing and Minimal Corrections for a Grammar Checker" In *Formal Grammar. 14th International Conference, FG 2009*.

Proceedings Article: Clément, Lionel, Kim Gerdes, and Renaud Marlet. "Grammaires d'erreur–correction grammaticale avec analyse profonde et proposition de corrections minimales" In *16e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2009)*, p. electronique. 2009.

Proceedings Article: Gerdes, Kim. "L'alignement pour les pauvres: Adapter la bonne métrique pour un algorithme dynamique de dilatation temporelle pour l'alignement sans ressources de corpus bilingues" *Proceedings of the 9es Journées internationales d'Analyse statistique des Données Textuelles* (2008a): 527-538.

Proceedings Article: Gerdes, Kim. "Poverty driven bilingual alignment" In *Proceedings of the International Symposium of Using Corpora in Contrastive and Translation Studies*. 2008b.

8.2.6 Other themes and book editions

Video: Gerdes, Kim. "L'exploration mécanique du possible : de la divination à l'écriture de toutes les idées" *Public debate with Michel de Broin in the Postdigital seminar on Machines organiques, machines textuelles.* École normale supérieure, Paris. 26/03/2018. https://www.youtube.com/watch?v=5I9YMwxOY6k

Edition: Gerdes, Kim, Eva Hajičová, and Leo Wanner (eds). *Dependency Linguistics: Recent advances in linguistic theory using dependency structures*. Vol. 215. Studies in Language Companion Series. John Benjamins Publishing Company, 2014.

Edition: Gerdes, Kim, Eva Hajičová, and Leo Wanner, eds. *Computational Dependency Theory*. Vol. 258. IOS Press, 2013.

Edition: Hajičová, Eva, Kim Gerdes, and Leo Wanner. *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*. 2013.

Edition: Gerdes, Kim, Eva Hajičová, and Leo Wanner. *Proceedings of the Dependency Linguistics 2011 conference (Depling 2011)*. 2011. Introduction.

Edition: Gerdes, Kim, David Beck, Jasmina Milićević, and Alain Polguère. *Proceedings of the 4th International Conference on Meaning-Text Theory*: 410 pages. 2009.

Edition: Gerdes, Kim, Tilmann Reuther, and Leo Wanner (eds.). *MTT 2007: Meaning-Text Theory 2007: Proceedings of the 3rd International Conference on Meaning-Text Theory*, Klagenfurt, May 20-24, 2007.

Journal Article: Gerdes, Kim. "Sur la non-équivalence des représentations syntaxiques: Comment la représentation en X-barre nous amène au concept du mouvement" *Cahiers de grammaire* 30: 175-192. 2006.

Edition: Gerdes, Kim, and Claude Muller. *Actes des 4èmes Journées de Syntaxe de l'ERSS : Ordre des mots dans la phrase française, positions et topologie*. *Linguisticae Investigationes*, 29-1. Introduction. 2006.

Thesis: Gerdes, Kim and Julia Kempe. *Décomposition équidimensionnelle de variétés en temps polynomial* Master Thesis. École polytechnique and University Paris XI. 1996.

9 Other References

Abbott, Ryan. "I Think, Therefore I Invent: Creative Computers and the Future of Patent Law." *BCL Rev.* 57 (2016): 1079.

Abbott, Ryan. “Everything is Obvious” *UCLA Law* Review, Vol 66(1) 2017. Forthcoming.

Abbott, Ryan, Jeremy Lack, and David Perkins. "Managing disputes in the life sciences." *Nature biotechnology* 36, no. 8 (2018): 697.

Adams, D. 1979. The hitchhiker’s guide to the galaxy. New York: Pocket Books.

Bar-Hillel, Yehoshua. “A quasi-arithmetical notation for syntactic description.” *Language* 29, no. 1 (1953): 47-58.

Baroni, M. and Bernardini, S., 2005. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, *21*(3), pp.259-274.

Barsky, Robert F. *Noam Chomsky: A Life of Dissent.* Cambridge, MA: MIT Press. 1997.

Bassett, Ben “Progress and Polytheism: Could an Ethical West Exist Without Christianity?” *Quillette.* 2018.

Becker, Tilman, Anne Kilger, Patrice Lopez, and Peter Poller. *Multilingual generation for translation in speech-to-speech dialogues and its realization in Verbmobil*. In Proceedings of ECAI 2000, Berlin, Germany, August 2000.

Bell, Colin, and Howard Newby, eds. *Doing sociological research*. Free Press, 1977.

Benzitoun, Christophe, Karën Fort, and Benoît Sagot. “TCOF-POS: un corpus libre de français parlé annoté en morphosyntaxe.” In *JEP-TALN 2012-Journées d’Études sur la Parole et conférence annuelle du Traitement Automatique des Langues Naturelles*, pp. 99-112. 2012.

Benzitoun, Christophe, Jeanne-Marie Debaisieux, and Henri-José Deulofeu. “Le projet ORFÉO: un corpus d’étude pour le français contemporain.” *Corpus* 15 (2016).

Berrendonner, Alain. “Pour une macro-syntaxe in Données orales et théorie linguistique.” *Travaux de linguistique* 21 (1990): 25-36.

Berwick, Robert C., and Amy S. Weinberg. “Parsing efficiency, computational complexity, and the evaluation of grammatical theories.” Linguistic Inquiry (1982): 165-191.

Buffier, Claude. *Grammaire française sur un plan nouveau, avec un Traité de la prononciation des e et un Abrégé des règles de la poésie française*, 1709. https://gallica.bnf.fr/ark:/12148/bpt6k50481x.pdf

Bizzocchi, A.L., 2017 How Many Phonemes Does the English Language Have? International Journal on Studies in English Language and Literature (IJSELL) Volume 5, Issue 10, October 2017, PP 36-46

Blanche-Benveniste, Claire, Bernard Borel, José Deulofeu, Jacques Durand, Alain Giacomi, Claude Loufrani, Boudjema Meziane, and Nelly Pazery. 1979. “Des grilles pour le français parlé.” *Recherches sur le Français Parlé* Aix-en-Provence, 163-206.

Blanche-Benveniste, Claire. Le Français Parlé: Etudes Grammaticales, CNRS, Paris. 1990.

Blanche-Benveniste, Claire. *Approches de la langue parlée*. Éditions Ophrys, 2010.

Bloomfield, Leonard. Language. Allen & Unwin, New York. 1933.

Boas, Franz. *General anthropology*. D.C. Heath and Company, 1938.

Bod, Rens. Beyond grammar: An experience-based theory of language. 1998.

Bod, Rens. A Comparative Framework for Studying the Histories of the Humanities and Science. *Isis*, *106*(2), pp.367-377. 2015.

Bohnet, Bernd, and Leo Wanner. 2010. Open Soucre Graph Transducer Interpreter and Grammar Development Environment. *Proceedings of LREC 2010*.

Borel, Émile. "Mécanique Statistique et Irréversibilité". *J. Phys.* (Paris). Series 5. 3: 189–196. 1913. https://hal.archives-ouvertes.fr/jpa-00241832/document

Borges, Jorge Luis. “A quien leyere”, *Preface to Fervor de Buenos Aires*. Imprenta Serantes. Buenos Aires. 1923.

Bouayad-Agha, Nadjet, Gerard Casamayor, Gabriela Ferraro, Simon Mille, Vanesa Vidal, and Leo Wanner. "Improving the comprehension of legal documentation: the case of patent claims." In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pp. 78-87. ACM, 2009.

Bühler, Karl, *Sprachtheorie. Die Darstellungsfunktion der Sprache*. G. Fischer, Jena. 1934.

Bußmann, Hadumod, and Hartmut Lauffer. *Lexikon der Sprachwissenschaft*. Stuttgart: Kröner. 2008.

Candito, Marie-Hélène, and Sylvain Kahane. "Can the TAG derivation tree represent a semantic graph? An answer in the light of Meaning-Text Theory." In *Proceedings of the Fourth International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+ 4)*, pp. 21-24. 1998.

Candito, Marie-Hélène. *Organisation modulaire et paramétrable de grammaires électroniques lexicalisées. Application au français et à l’italien*. PhD thesis, Université Paris 7. 1999.

Clark, Alexander, and Shalom Lappin. *Linguistic Nativism and the Poverty of the Stimulus*. John Wiley & Sons, 2010.

Clarke, Arthur Charles. *The Nine Billion Names of God*. Harcourt, 1967.

Chomsky, Noam. *Aspects of the Theory of Syntax*. MIT Press, 1965.

Chomsky, Noam. *New Horizons in the Study of Language and Mind,* Cambridge, England, Cambridge University Press. 2000.

Coch José. “Overview of AlethGen”, Proc. 8th Int. Workshop on Natural Language Generation (INLG’96), Vol. 2, Herstmonceux, 25-28. 1996.

Coch, J. (1998). Interactive generation and knowledge administration in multimeteo. Workshop on Natural Language Generation (INLG98).

Cristianini, N., 2014. On the current paradigm in artificial intelligence. *AI Communications*, *27*(1), pp.37-43. https://pdfs.semanticscholar.org/a18e/c6116c69e5a559e2271af0add5424c7217be.pdf

Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press.

Croft, William, and D. Alan Cruse. *Cognitive linguistics*. Cambridge University Press, 2004.

Damourette, Jacques, and Édouard Pichon. *Des mots à la pensée. Essai de grammaire de la langue française. Paris: d’Artrey*. 1911-1940.

Dawkins, Richard. “The selfish gene: with a new introduction by the author.” *UK: Oxford University Press.* Originally published in 1976. 2006.

Dawkins, Richard. *The God Delusion*. Random House. 2006.

Demberg, Vera, *Syntactic processing in humans: time course, shallow processing and processing failure*, Invited Talk, IWPT 2017.

Descartes, René. *Discours de la méthode.* 1637. https://fr.wikisource.org/wiki/Discours_de_la_méthode

Desprez-Bouanchaud, Annie, Janet Doolaege, Lydia Ruprecht, and Breda Pavlic-UNESCO. “Guidelines on gender-neutral language.” (1999).

Devitt, Michael. *Ignorance of language*. Oxford University Press, 2006.

Dreyfus, H.L., *What computers still can’t do: A critique of artificial reason*. MIT press. 1992.

Eco, Umberto. *Kant and the platypus: Essays on language and cognition*. HMH. 2000.

Ehrenfeld, David W. *The arrogance of humanism*. Oxford University Press, 1981.

Duchier, Denys, and Ralph Debusmann. "Topological dependency trees: A constraint-based account of linear precedence." In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pp. 180-187. Association for Computational Linguistics, 2001.

Fillmore, Charles J., Paul Kay, and Mary Catherine O’connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language.* 501-538.

Fisher, Ronald A. “The goodness of fit of regression formulae, and the distribution of regression coefficients.” *Journal of the Royal Statistical Society* 85, no. 4: 597-612. 1922.

Foucault, Michel. *The order of things*. Routledge, 2005. Original: Michel, Foucault. “Les mots et les choses.” Paris, Gallimard (1966).

Gabelentz, Georg von der. *Die Sprachwissenschaft, ihre Aufgaben, Methoden und bisherigen Ergebnisse.* 1891. Republished by Language Science Press, Classics. 2016. http://langsci-press.org/catalog/book/97

Gaiffe, Bertrand, Benoît Crabbé, and Azim Roussanaly. “A new metagrammar compiler.” In *Proceedings of the Sixth International Workshop on Tree Adjoining Grammar and Related Frameworks (TAG+ 6)*, pp. 234-241. 2002.

Gibson, Edward. "Linguistic complexity: Locality of syntactic dependencies." *Cognition* 68, no. 1 (1998): 1-76.

Girard, Gabriel. *Les vrais principes de la langue francoise; ou, La parole, reduite en methode conformement aux loix de l’usage.* 1747. https://gallica.bnf.fr/ark:/12148/bpt6k50629c.pdf

Gleason Jr, Henry A. *An introduction to descriptive linguistics*. Holt, New York. 1955.

Golumbia, D., 2004. The interpretation of nonconfigurationality. *Language & Communication*, *24*(1), pp.1-22.

Gore, R. J., Lemos, C., Shults, F. L., & Wildman, W. J. (2018). Forecasting Changes in Religiosity and Existential Security with an Agent-Based Model. *Journal of Artificial Societies and Social Simulation*, *21*(1).

Graffi, G. (2001). *200 Years of Syntax: A critical survey* (Vol. 98). John Benjamins Publishing.

Gross, Maurice. "On the failure of generative grammar." *Language* (1979): 859-885.

Hajičová, E., 2011. *Computational linguistics without linguistics? A view from Prague.* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.369.7027&rep=rep1&type=pdf

Halevy, A., Norvig, P. and Pereira, F., 2009. The unreasonable effectiveness of data. IEEE Intelligent Systems, 24(2), pp.8-12.

Halliday, Michael Alexander Kirkwood. "Categories of the theory of grammar." *Word* 17, no. 2: 241-292. 1961.

Halliday, Michael Alexander Kirkwood. *On Grammar*, Bloomsbury Publishing. 2005.

Harari, Yuval Noah. *Homo Deus: A brief history of tomorrow*. Random House, 2016.

Harris, Sam. *The moral landscape: How science can determine human values*. Simon and Schuster, 2011.

Hattenbach, Ben, and Joshua Glucoft. "Patents in an Era of Infinite Monkeys and Artificial Intelligence." *Stanford Technology Law Review.* 19 (2015): 32.

Hernandez, Madeleine, *Automatic Generation of Hip-Hop and Rap Lyrics, Master Thesis*, Sorbonne Nouvelle. 2018.

Hockett, Charles Francis. *A manual of phonology*. No. 11. Waverly Press, 1955.

Höhle, Tilman, and Albrecht Schöne. "Der Begriff „Mittelfeld “, Anmerkungen über die Theorie der topologischen Felder." *Beiträge zur deutschen Grammatik* (1986): 279.

Honnibal, M. and Johnson, M., 2014. Joint incremental disfluency detection and dependency parsing. *Transactions of the Association of Computational Linguistics*, *2*(1), pp.131-142.

Isabelle, P. and Kuhn, R., 2018. A Challenge Set for French--> English Machine Translation. *arXiv preprint arXiv:1806.02725*.

Joshi, Aravind K. “How much context-sensitivity is necessary for characterizing structural descriptions – Tree Adjommg Grammars” In *Natural Language Processing – Theoretical, Computational and psychological Perspectives*, Cambridge Umversity Press. 1985.

Joshi, Aravind K., Tilman Becker, and Owen Rambow. Complexity of scrambling: A new twist to the competence/performance distinction. Appeared in 3e Colloque International sur les grammaires d’Arbres Adjoints (TAG+3). 1994.

Humboldt, Wilhelm von. “Über den Geschlechtsunterschied und dessen Einfluß auf die organische Natur.” *Werke in fünf Bänden*. Eds. Andreas Flitner and Klaus Giel. 1795. 268-295. https://www.friedrich-schiller-archiv.de/die-horen/die-horen-1795-stueck-2/v-ueber-geschlechtsunterschied-und-einfluss-auf-organische-natur/

Humboldt, Wilhelm von (1836) *Über die Verschiedenheitdes menschlichen Sprachbaues und ihren Einfluss auf die geistige Entwickelung des Menschengeschlechts*. Translated as *‘On language’: On the diversity of human language construction and its influence on the mental development of the human species*. Cambridge University Press, 1988.

Isabelle, P. and Kuhn, R., 2018. A Challenge Set for French --> English Machine Translation. *arXiv preprint arXiv:1806.02725*.

Jakobson, Roman. "On linguistic aspects of translation." *On translation*, Harvard University Press. 1959.

Kaplan, Ronald M., and Joan Bresnan. "Lexical-functional grammar: A formal system for grammatical representation." *Formal Issues in Lexical-Functional Grammar* 47 (1982): 29-130.

Kaplan, Ronald M., and Annie Zaenen. "Long-distance dependencies, constituent structure, and functional uncertainty." *Formal Issues in Lexical-Functional Grammar* 47 (1995): 137-165.

Kathol, Andreas. "Linearization-based German syntax." PhD diss., The Ohio State University, 1995.

Kathol, Andreas. *Linear syntax*. Oxford University Press, USA, 2000.

Kirilin, Angelika and Yannick Versley. 2015. What is hard in Universal Dependency Parsing. *Proceedings of the 6th Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2015)*.

Kahane, Sylvain, and Paola Pietrandrea. “La typologie des entassements en français.” In *SHS Web of Conferences*, vol. 1, pp. 1809-1828. EDP Sciences, 2012.

Kahane, Sylvain. *How dependency syntax appeared in the French Encyclopedia: from Buffier (1709) to Beauzée (1765)*. To appear.

Klasios, J. “Cognitive traits as sexually selected fitness indicators” Review of General Psychology. pp. 428–442. 2013.

Kuhn, Thomas S. *The structure of scientific revolutions*. University of Chicago press, 1962.

Lakatos, Imre. “Criticism and the methodology of scientific research programmes.” In *Proceedings of the Aristotelian society*, vol. 69, pp. 149-186. Aristotelian Society, Wiley, 1968.

Lebart, L. and Salem, A. *Statistique textuelle*. Paris: Dunod. 1994.

Lee, Timothy B. “How Democrats and trial lawyers killed patent reform” *Wired.* 2014. https://www.vox.com/2014/5/23/5742620/patent-reform-is-dead-heres-who-killed-it

Leech, Geoffrey. “100 million words of English the British National Corpus (BNC)” *English Today* 9, no. 1: 9-15. 1993.

Lenat, D.B. and Guha, R.V. *Building large knowledge-based systems; representation and inference in the Cyc project.* Addison-Wesley Longman Publishing Co. 1989.

Léon, J. “Empiricism versus Rationalism revisited. Current Corpus Linguistics and Chomsky’s arguments against corpus, statistics and probabilities in the 1950-1960s”. In S. Matteos & P. Schmitter (eds.) *Linguistische und epistemologische Konzepte*. Münster: Nodus Publikationen:157 – 176. 2007. http://htl.linguist.univ-paris-diderot.fr/_media/leon/empiricism2007.pdf

Lessig, Lawrence. *The future of ideas: The fate of the commons in a connected world*. Vintage, 2002.

Li, Yixuan, *Development of an Adapted Method of Word Segmentation for Chinese Technological Texts, Master Thesis*, Sorbonne Nouvelle. 2018.

Liu, Haitao. “Dependency distance as a metric of language comprehension difficulty.” *Journal of Cognitive Science*, 9. 2 (2008): 159-191.

Losonsky, Michael, “Introduction”. In Humboldt: *On Language: On the Diversity of Human Language Construction and its Influence on the Mental Development of the Human Species*. Cambridge Texts in the History of Philosophy. 2nd Edition. 1999.

MacWhinney, B. and Snow, C. The child language data exchange system. Journal of child language, 12(2), pp.271-295. 1985.

McCowan, Brenda, Hanser Sean F., Doyle Laurence R. Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. *Anim. Behav. 57*:409–19. 1999.

McLaughlin, Michael. “Computer-Generated Inventions.” Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3097822. 2018.

Mel’čuk, Igor Aleksandrovič. *Dependency syntax: Theory and Practice*. SUNY Press, 1988.

Mel’čuk, Igor Aleksandrovič. “Paraphrase et lexique dans la théorie linguistique Sens-Texte in Lexique et paraphrase.” *Lexique* 6 (1988): 13-54. 1988b.

Mel’čuk, Igor Aleksandrovič. “Collocations and lexical functions.” *Phraseology. Theory, analysis, and applications* (1998): 23-53.

Mikolov, Thomas, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781. 2013.

Miller, George A. "The magical number seven, plus or minus two: Some limits on our capacity for processing information." *Psychological review* 63, no. 2 (1956): 81.

Moore, Peter. The varied ways plants tap the Sun. *New Scientist*, 81 (February 12), 394-397. 1981.

Moscato, Pablo. "On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms." *Caltech concurrent computation program, C3P Report* 826 (1989): 198

Müller, Stefan. *Deutsche Syntax deklarativ: head-driven phrase structure grammar für das Deutsche*. Vol. 394. Walter de Gruyter GmbH & Co KG, 1999.

Nivre, Joakim, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan T. McDonald et al. "Universal Dependencies v1: A Multilingual Treebank Collection." In *LREC*. 2016.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S. and Yuret, D. “The CoNLL 2007 shared task on dependency parsing” In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. 2007.

Osborne, Timothy J. “Tests for constituents: What they really reveal about the nature of syntactic structure.” *Language Under Discussion* 5, no. 1 (2018): 1-41.

Oostdijk, Nelleke, Eva D'hondt, Hans Van Halteren, and Suzan Verberne. "Genre and domain in patent texts." In *Proceedings of the 3rd international workshop on Patent information retrieval*, pp. 39-46. ACM, 2010.

Paul, Hermann, *Prinzipien der Sprachgeschichte*. Halle: Max Niemeyer, 1st ed. 1880; 3d ed. 1898.

Payne, Geoff Surveys, Statisticians and Sociology: A History of (a Lack of) Quantitative Methods, Enhancing Learning in the Social Sciences, 6:2, 74-89. 2014.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L., 2018. Deep contextualized word representations. *Proceedings of NAACL, arXiv preprint arXiv:1802.05365*.

Péroz, P., 2009. On ne dit pas ouais!. *Langue française*, (1), pp.115-134.

Pietrandrea, Paola, Sylvain Kahane, Anne Lacheret-Dujour, and Frédéric Sabio. “The notion of sentence and other discourse units in corpus annotation” In *Spoken Corpora and Linguistic Studies*. John Benjamins, 2014.

Pinker, Steven. *How the mind works*. Penguin books, 1999.

Plotkin, Robert, *The genie in the machine*. Stanford University Press. 2009.

Poibeau, Thierry. *Traitement automatique du contenu textuel*. Lavoisier, 2011.

Polguère, Alain. "La théorie sens-texte." *Université de Montréal URL: http://olst.ling.umontreal.ca/pdf-apol/PolgIntroTST.pdf* (1998).

Postal, P.M., 2004. The openness of natural languages. *Skeptical linguistic essays*, pp.173-201.

Pullum, Geoffrey K. and Barbara C. Scholz (2001) “On the Distinction between Model-Theoretic and Generative-Enumerative Syntactic Frameworks”, in LACL 2001, Proceedings of the Conference on Logical Aspects of Computational Linguistics, Berlin, Springer-Verlag.

Pullum, Geoffrey K., and Barbara C. Scholz. “Empirical assessment of stimulus poverty arguments.” *The linguistic review* 18.1-2 (2002): 9-50.

Radick, Gregory. “The unmaking of a modern synthesis: Noam Chomsky, Charles Hockett, and the politics of behaviorism, 1955–1965.” *Isis* 107.1 (2016): 49-73.

Rambow, Owen, and Aravind Joshi. 1994. A formal look at dependency grammars and phrase-structure grammars, with special consideration of word-order phenomena. *Recent trends in Meaning-Text theory* 39 (ed. Leo Wanner): 167-190.

Rehbein, Ines, Julius Steen, and Bich-Ngoc Do. 2017. Universal Dependencies are hard to parse–or are they? *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*.

Russell, Bertrand. *Why I am not a Christian*. Watts, 1927.

Saussure, Ferdinand de. *Cours de linguistique générale*, Payot 1916.

Roy Schwartz, Omri Abend, Ari Rappoport. 2012. Learnability-based syntactic annotation design. In *Proceedings of COLING 24*, 2405–2422.

Shim, Kyungmin. “소수 언어쌍 병렬 코퍼스 구축 방안:영어자막 활용을 중심으로″ (Building a parallel corpus from movie subtitles for rare language pairs). *The Korean Generative Grammar Circle*. Proceedings of KGGC 2017 Spring Joint Conference. 2017.

Sokal, A.D. and Bricmont, J., 1998. *Intellectual impostures: postmodern philosophers’ abuse of science*. London: Profile books.

Schuurman I., W. Goedertier, H. Hoekstra H., N. Oostdijk, R. Piepenbrock, M. Schouppe (2004), Linguistic annotation of the Spoken Dutch Corpus: If we had to do it all over again ..., in Proceedings of LREC, Lisbon, 57-60.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. “Dropout: a simple way to prevent neural networks from overfitting.” *The Journal of Machine Learning Research* 15, no. 1: 1929-1958. 2014.

Sczesny, Sabine, Magda Formanowicz, and Franziska Moser. “Can gender-fair language reduce gender stereotyping and discrimination?.” *Frontiers in psychology* 7 (2016): 25. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4735429/

Silveira, Natalia and Christopher Manning. 2015. Does Universal Dependencies need a parsing representation? An investigation of English. *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*.

Skinner, Burrhus Frederic. *Verbal behavior*. Appleton-Century-Croft, 1957.

Summers-Stay, Douglas. *Machinamenta: The Thousand Year Quest to Build a Creative Machine*. CreateSpace Independent Publishing Platform. 2011

Trabant, Jürgen. *Weltansichten: Wilhelm von Humboldts Sprachprojekt*. CH Beck, 2012.

Thompson, Sandra A., and William C. Mann. “Rhetorical structure theory.” *IPRA Papers in Pragmatics* 1, no. 1 (1987): 79-105.

Vallduví, Enric, 1995. Structural properties of information packaging in Catalan. *Discourse configurational languages*, pp.122-152.

Vaugelas, Claude Favre de. *Remarques sur la langue françoise utiles à ceux qui veulent bien parler et bien escrire*. https://fr.wikisource.org/wiki/Remarques_sur_la_langue_française,_utiles_à_ceux_qui_veulent_bien_parler_et_bien_écrire “1647.

Voirin, Bryson, Roland Kays, Martin Wikelski, and Margaret Lowman. "Why Do Sloths Poop on the Ground?." In *Treetops at Risk*, pp. 195-199. Springer, New York, NY, 2013.

Wanner, Leo, Sören Brügmann, Boubacar Diallo, Mark Giereth, Yiannis Kompatsiaris, Emanuele Pianta, Gautam Rao, Pia Schoester, and Vasiliki Zervaki. "PATExpert: Semantic Processing of Patent Documentation." *Proceedings of Semantics and Digital Media Technologies*. 2006.

Wanner, Leo, Bernd Bohnet, Nadjet Bouayad-Agha, François Lareau, and Daniel Nicklaß. 2010. MARQUIS: Generation of user-tailored multilingual air quality bulletins. Applied Artificial Intelligence 24, no. 10: 914-952.

Weigel, Peter. Aquinas on Simplicity: An Investigation into the Foundations of His Philosophical Theology. Bern: Peter Lang, 2008.

Weir, David Jeremy. 1988. Characterizing mildly context-sensitive grammar formalisms. PhD thesis, University of Pennsylvania.

Winter, Yoad. “A unified semantic treatment of singular NP coordination.” *Linguistics and philosophy* 19, no. 4 (1996): 337-391.

Wittgenstein, Ludwig. *Das Blaue Buch: Eine Philosophische Betrachtung; Zettel*. Vol. 5. M., Suhrkamp, 1970. Notes dictated 1933–1934.

Wolf, L. (1983). La normalisation du langage en France. De Malherbe à Grevisse. *La Norme linguistique, Québec/Paris: Gouvernement du Québec, Conseil de la langue française, Le Robert*, 105-137.

Yu, J., Elkaref, M. and Bohnet, B., 2015. Domain adaptation for dependency parsing via self-training. In *Proceedings of the 14th International Conference on Parsing Technologies* (pp. 1-10).

Zimina, Maria. “Alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles.” *Actes des 7es Journées internationales d’Analyse statistique des Données Textuelles*: 1195-1202. 2004.

Zeman, D. and Philipp Resnik. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*. 2008.