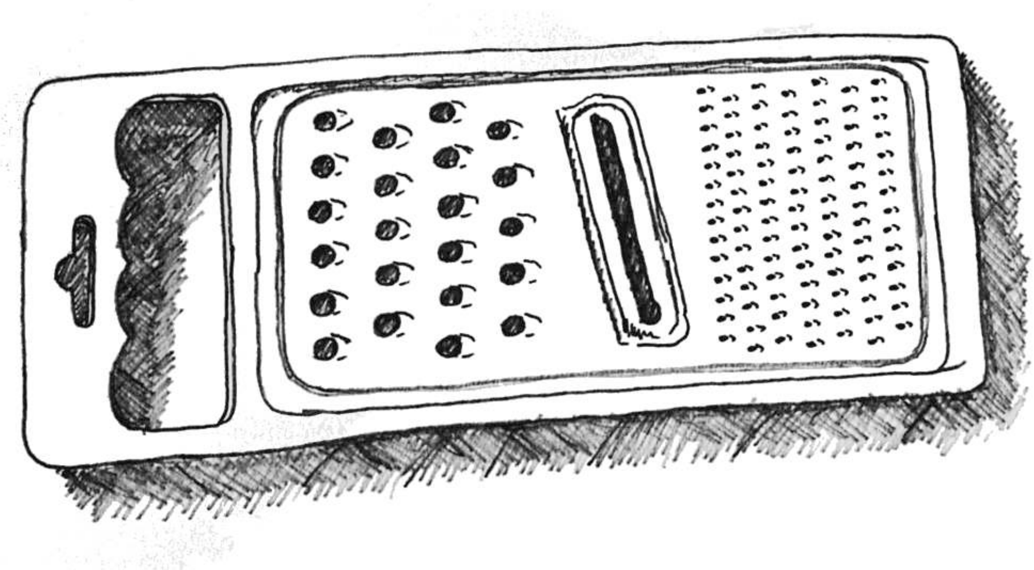
The Limits of Machine Translation



Mathias Winther Madsen

Thesis submitted for the degree of
Master in Information Technology and Cognition

Department of Scandinavian Studies and Linguistics,
Faculty of Humanities, University of Copenhagen

Supervisor: Bolette Sandford Pedersen,
Center for Language Technology, University of Copenhagen

December 23, 2009

Contents

Danish abstract

Dette speciale handler om maskinoversættelse, det vil sige automatisk oversættelse fra ét naturligt sprog til et andet ved hjælp af computere. Sådanne oversættelser er ofte præget af meningsforstyrrende fejl, hvilket jeg giver eksempler på. Specialets mål er at analysere hvorfor samt at drage nogle konklusioner omkring maskinoversættelsens fremtid på baggrund af denne analyse.

Konklusionen på min analyse er, at en væsentlig del af fejlene i maskinoversættelser skyldes den interaktion, der er mellem menneskers hverdagsliv og deres sprog. Denne konklusion er dels baseret på en historisk analyse af de problemer, som maskinoversættelse indtil nu er stødt på (kapitel 2), og dels på nogle mere generelle argumenter omkring sprogets rolle i livet (kapitel 3).

Disse mere generelle argumenter bygger på pointer hentet hos filosofferne Ludwig Wittgenstein og Martin Heidegger. Jeg giver imidlertid også nogle konkrete eksempler på sproglige fænomener, som viser, hvordan denne interaktion mellem livet og sproget kommer til udtryk (kapitel 4).

En central del af min forklaring af problemerne med maskinoversættelse er, at den praktiske erfaring, der skal til for at oversætte en tekst, ikke lader sig beskrive i et formelt system. På den baggrund konkluderer jeg, at oversættelse udført udelukkende ved hjælp af computere af principielle årsager er forhindret i at nå samme kvalitetsniveau som oversættelse udført helt eller delvist af mennesker.

Acknowledgements

I owe a special thanks to Martin Stokhof of the Institute for Logic, Language, and Computation at the University of Amsterdam for his help with this thesis throughout the spring of 2009.

Without him, this text would certainly have been possible, but not as much fun writing.

Chapter 1

Introduction

1.1 The topic and scope of this text

Do you see the difference between the use of the word *executed* in the following sentences?

(1.1) (a) The command interface defines a single method called "execute" that is invoked by the internal CommandExecutor when a command is to be executed.

(b) An Iranian cleric, Hojatoleslam Rahimian, called today for the leaders of Iran's opposition Green Movement to be executed.

If you do, you possess a skill that hardly any computer programme does. Software for automatic translation or "machine translation" is notorious for missing such distinctions, and machine translations are often poor in quality or downright comical.

Since machine translation has been around for almost six decades by now, it is natural to ask why we are not doing better. In this thesis, I propose an explanation of that fact which frames machine translation in a broader theory of human nature and conduct. This enables me to outline the limits of machine translation and explain why it seems to have run repeatedly against a brick wall in the course of its history.

This text is thus not a thesis "in" machine translation in the sense that it will bring us closer to the ultimate goal of "fully automatic, high-quality machine translation" (Bar-Hillel, 1964a). In fact, the conclusion of my analysis is that many of the problems that machine translation faces cannot be solved at all. High-quality translation done solely by machines is not possible and machine translated texts will continue to be plagued by errors in the future, ranging from eccentric turns of phrase to grave distortions of meaning.

This does not mean that machine translation is useless or that computers have no role to play in translation. A low-quality machine translation is better than no translation, and good software may help a translator do her job faster and better. Machine translation is not impossible in the sense that a computer cannot produce any kind of output, or in some cases produce fairly good output.

But there is a difference between genuinely competent translation and what the machine translation researcher Victor H. Yngve denounced as "the 95% approach" (Yngve, 1957, p. 60). When you translate a law, a job application, a fire emergency instruction, a military order, or a medical prescription, you do not want the translation to be "fairly clear" and "almost accurate," but clear and accurate. As two more recent researchers dryly commented, "a 95% system in the worst case produces a translated text analogous to a jar of cookies, only 5% of which are poisoned" (Carbonell and Tomita, 1987, p. 69). There are plenty of actual practical situations where a 5% error rate is outrageously high. Human beings are generally not prone to confuse the execution of a command with the execution of a political leader, and we expect our translations to reflect that.

The question is whether computers are lagging behind in principle and by their very nature, or whether we can fix the problems. Maybe we just need to equip our computers with some statistics, so they know that *will* is more likely to a verb than a noun. Maybe they just need knowledge bases to tell them that there are no pickled body parts in the sentence *I closed the jar with my hands*. Or maybe they just need to look at some human translations and figure out the principles themselves. And maybe then, they will eventually translate as well as we do.

Such hopes are not unnatural, but I do believe that they are unwarranted. This belief is based on a general theory of linguistic practice that I have borrowed from the philosophers Ludwig Wittgenstein and Martin Heidegger. Briefly put, they argue that formal models of language fail to recognise the sources of linguistic meaning. The bare bones of this argument are presented in chapter 3.

The other chapters serve to relate this theory to the practice of machine translation. Chapter 2 contains a historical introduction to the techniques used for machine translation and includes some preliminary reflections on where and why they fail. In chapter 4, I relate the theory of language from chapter 3 to some concrete linguistic phenomena. The remainder of this chapter, I will spend introducing the basic components of the discussion.

1.2 The "disciplined but unintelligent" clerk

Before I turn to the more specific issue of computers in translation, I want to make some points about computers in general. Briefly put, I want to argue that computer models always come with theoretical commitments. A necessary part of a computer solution to a problem is a conception of what the problem was in the first place.

The British logician Alan Turing, who was intimately involved in the construction of the first computers, was well aware of this fact. He repeatedly urged future programmers to think properly about machines, as in his manual for the Mark II computer in Manchester:

> Electronic computers are intended to carry out any definite rule of thumb process which could have been done by a human operator working in a disciplined but unintelligent manner. The electronic computer should however obtain its results very much more quickly. (Turing, 1951, p. 1)

These "rules of thumb" were to be written by a programmer and then implemented on the machine. In his proposal for what was later realised as the ACE computer, Turing explains how this division of labour was meant to work:

> It is intended that the setting up of the machine for new problems shall be virtually only a matter of paper work. Besides the paper work nothing will have to be done except prepare a pack of Hollerith cards in accordance with the paper work, and to pass them through a card reader connected with the machine. (Turing, 1946, p. 3)

The real work is thus done in the "paper" process of rewriting the task in a machine format. This human preprocessing of the problem is the same that we find in any programming process, be it by punch card and soldering iron or by high-level programming languages.

The features of this process is also not dependent on programming scheme or machine architecture. Connectionist networks work within a given set of rules and are trained on a preselected set of input–output pairs, tailored for the specific program (cf. for instance Rumelhart, 1989). Programs applying statistical methods need both a preselected corpus and hand-written rules for applying the statistics extracted from the corpus (cf. section 2.3). We should not push the metaphors of "fuzzy," "neural," or "trained" systems too far. Many programming schemes allow for automatic improvement, but only when the programmer selects data and the "better" responses in advance. That is, put in more human-centred terms, the machine's only contact with social and practical life comes through the generalisations provided by the programmer.

Solving a problem by means of a computer thus comes down to analysing it as a series of primitive processes. A problem is easy if it can easily be translated into easy problems (cf. Hromkovič, 2004, ch. 5.3). We are already familiar with reductions from everyday life. For instance, the multiplication algorithm we learn in school reduces multiplication of multi-digit numbers to a series of additions inspections of a table of the products 1×1 through 9×9.

Some problems, however, cannot be translated into a given set of routines in a computationally cheap way. Others cannot be translated at all. There is, for instance, no general algorithm for how to find the trisection point of a given angle by drawing only circles and straight line segments. There are specific solutions to specific cases, but no general method that covers all cases.

In his 1949 memo to the Rockefeller Foundation (reprinted as Weaver, 1955), Warren Weaver wondered whether translation was a problem that could be analysed in simple, arithmetic terms:

> [I]t is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the "Chinese code." If we have useful methods for solving almost any cryptographic problem, may it not be that with proper interpretation we already have useful methods for translation? (Weaver, 1955, p. 22)

The claim in this thesis is that we do not have such methods. However, the problems do not stem from computational complexity or logical undecidability, and you will not find time complexity analyses or Gödel numberings in this text. The problem I want to point out is that the everyday experience of being a human being is necessary for ordinary translation.

The philosopher Robert Brandom (2008, Lecture One, p. 11ff) has made some points about explications of human practice that may help to pave the way for this argument. In his model, the correct application of a vocabulary V requires a set of practices-or-abilities P. Underlying any linguistic competence is thus a relation of "PV-sufficiency" (Brandom, 2008, Lecture One, p. 11).

However, this practice may itself be open to explication—we can talk about talking. In some situations, we may even be able to teach someone a certain sublanguage, like the one containing statements like "$16\times16 = 256$," by explicit rules. If this explication is non-trivial, the instructions are given in a different vocabulary V'. This, in its turn, is powered by its own set of practices-or-abilities P' (cf. figure 1.1).

A manual for a vocabulary V can thus be thought of as a way to link between two practices, P and P'. The practices-or-abilities in P' are left unanalysed and may in principle include anything from multiplication tables to divine inspiration. The manual thus represents a shift from one unanalysed remainder to another.

Translation does not require any mysterious, magical component, but it does require a kind of experience that machines do not have. The "system" consisting of a person with a body, a brain, a past, and an environment has one particular set of primitive practices-or-abilities. A machine has a different one. In the words of the artificial intelligence researcher Terry Winograd:

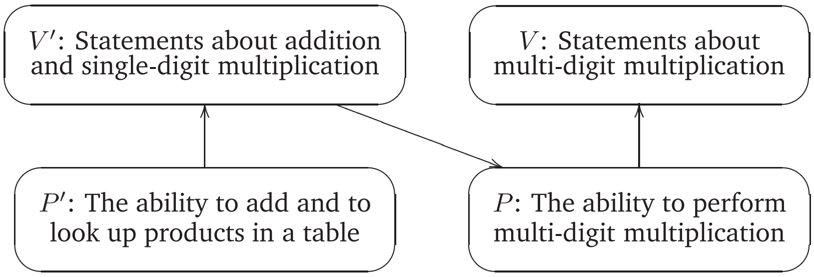


Figure 1.1: An analysis of one linguistic practice in terms of another, based on Brandom (2008).

> Indeed, artificial intelligence has not achieved creativity, insight and judgement. But its shortcomings are far more mundane: we have not yet been able to construct a machine with even a modicum of common sense or one that can converse on everyday topics in ordinary language.
>
> The source of the difficulties will not be found in the details of silicon micro-circuits or of Boolean logic. The basic philosophy that has guided the research is shallow and inadequate, and has not received sufficient scrutiny. (Winograd, 1991, p. 198–99)

Explaining the deficiency in terms of some romantic conception like the soul is not an option, even though some researchers have moved in that direction (Weizenbaum and Searle, for instance). The problem is not that we do not know how a neuron works or that we cannot simulate it. The problem is that the experience causing our neurons to connect in one way rather than another is far too rich and diverse to account for in a table of frequencies or production rules.

If this claim is true, we should expect a recurring pattern in all attempts to decompose the translation process: Whenever we invent some ingenious algorithm which breaks up the problem into smaller subproblems, we will find on closer scrutiny that at least one of the subproblems cannot be solved. In most cases, such sticky issues are dumped in a box labelled "contextual effects." We may try to account for this "context," but that will just produce yet another formalisation with its own "contextual" exceptions. It will leave it to another subroutine, another module, another department, another field, or another time to sort out the remaining problems.

Hubert Dreyfus and Charles Spinosa make the same point while discussing the context-sensitiveness of apparently simple sentences like *The cat is on the mat*:

> [W]hatever new features—such as being in a nontheatrical context—the essentialist introduces to bring the deviant cases under a refined account of the type would be subject to the same open-endedness of interpretation as the features originally defining the type. So if the essentialist adds "in a nontheatrical context" to his account of the cat-on-the-mat type, the antiessentialist will invent a context where it is unclear what counts as a nontheatrical context. (Spinosa and Dreyfus, 1996, p. 740)

So the more rigid our formal categories are, the more informal they will have to be applied in order to account for everything. An example of this phenomenon appears in Victor Yngve's 1957 list of six types of contextual clues for disambiguation. He concludes the list with the item

> 6) All other contextual clues, especially those concerned with an exact knowledge of the subject under discussion. These will undoubtedly remain the last to be mechanized.
>
> Finding out how to use these clues to provide correct and accurate translations by machine presents perhaps the most formidable task that language scholars have ever faced. (Yngve, 1957, p. 60)

Every single theory I am going to discuss in the following sections include such "everything else" categories, which are left to be analysed further in a near future.

Formal linguistics has produced some enlightening and powerful theories indeed, but we should always ask ourselves what kind of practices-or-abilities we would need to apply them appropriately. If a theory has just a single exception, it is fair to ask where the exception-handling takes place. My claim here is that the exceptions always wash ashore at the same place, namely at the feet of the human reader and his or experience of the world.

1.3 The future of machine translation

The case made in this thesis—that machine translation will continue to produce errors of arbitrary severity—has several predecessors, most notably Yehoshua Bar-Hillel, Hubert Dreyfus, Terry Winograd, Fernando Flores, and Martin Kay (Bar-Hillel, 1964a; Dreyfus, 1965; Dreyfus et al., 1986; Dreyfus, 1992, 2007; Winograd, 1991; Winograd and Flores, 2005; Kay, 1980, 1997; Robinson, 1991).

It is, however, far from consensus in the machine translation research community. In this section, I will, for the sake of contrast, cite some examples of the kind of rhetoric one finds in the popular and technical literature about machine translation.

1.3.1 Looking ahead from 2009

In an essay named "The end of translation as we know it," the president of Byte Level Research, John Yunker, predicts that statistical machine translation "may eliminate the need for 'commodity' translation" (Yunker, 2008, p. 31). He continues with yet another prediction:

> Imagine a future generation of Google's Chrome browser with SMT technology built right in, so when you visited a website in a different language, Chrome automatically inserted the translation. It's not going to happen anytime soon, but I think it will happen. (Yunker, 2008, p. 31)

We can find a similar claim in the German internet journal *Computerwoche*:

> The human dream of a universal language is almost as old as mankind itself. With statistical machine translation, this dream seems to be achievable.[1] (Hülsbörner, 2007, unpaginated)

The article then quotes Philipp Koehn from the machine translation group at the University of Edinburgh as saying that "we can translate technical and political texts with no problems whatsoever"[2] (Hülsbörner, 2007, unpaginated).

[1] "Der menschliche Traum von der Universalsprache ist fast so alt wie der Mensch selbst. Mit statistischen maschinellen Übersetzungen scheint er greifbar zu werden."

[2] "Technische und politische Texte können wir völlig problemlos übersetzen lassen."

In the European Commission, the optimism is slightly more modest: “In terms of quality, machine translation will remain inferior to human translation for many years,” one recent report states (Accipio Consulting, 2006, p. 25). However, the insufficiencies will gradually be overcome, we are told, and machine translation services will gradually be able to move to tasks with higher quality requirements:

> Human translation will prevail in all areas where high quality is an absolute necessity. In contrast, machine translation will take over the low end of the market . . . With quality and performance improvement over time, machine translation will move up-market. (Accipio Consulting, 2006, p. 25)

Such prophecies of what is to come in a near but not too near future are not uncommon in the commercial world. One 2007 flyer from the Danish Business Research Academy hints in the same direction:

> Can one make a machine translate the collected works of Shakespeare into Danish? – Maybe not yet. But within with regard to technical texts, the technology has made great progress within the last few years.[3] (Danmarks ErhvervsforskningsAkademi)

Unsurprisingly, the Google corporation also gives an good assessment of their own translator, asserting that they have “achieved very good results in research evaluations” (Google, 2009). But they too recognise that it may take “some time” before machine translation reach the quality of standard translation:

> Even today’s most sophisticated software, however, doesn’t approach the fluency of a native speaker or possess the skill of a professional translator. Automatic translation is very difficult, as the meaning of words depends on the context in which they’re used. While we are working on the problem, it may be some time before anyone can offer human quality translations. (Google, 2009)

At the industry standards organisation LISA, the same pattern reoccurs—problems are recognised, but we should be able to overcome them soon with a little extra money:

> A close look at research into machine translation shows that very high quality systems are within reach, but developing them for freely variable texts requires more funding than is available today. (Dillinger and Rommel, 2004, p. 45)

So not everything is possible today, but we seem to be at a critical moment in time—with automatic translation just within grasp a couple of years ahead of us.

1.3.2 Looking ahead from 1949

As I mentioned on page 7, the success of cryptography during the second world war had given Warren Weaver a sense that translation was just another mathematical problem waiting to be solved (cf. Weaver, 1955, pp. 18 and 22). By 1955, his optimism had not diminished. In the foreword to *Machine Translation of Languages* (Locke and Booth, 1955) he reported:

[3]“Kan man få en maskine til at oversætte Shakespeares samlede værker til dansk? – Måske ikke endnu. Men inden for oversættelse af tekniske tekster har teknologien i de sidste år taget stormskridt.”

> Students of languages and of structures of languages, the logicians who design computers, the electronic engineers who build and run them—and specially the rare individuals who share all of these talents and insights—are now engaged in erecting a new Tower of Anti-Babel . . . [that they hope] will build part of the way back to that mythical situation of simplicity and power when men could communicate freely together, and when this contributed so notably to their effectiveness. (p. vii)

Like weaver, the British engineer Andrew Donald Booth and his collegue Bill Locke seemed to think that machine translation was just around the corner:

> It is perhaps not going to far to state that the main lines of the principles on which a mechanical translation depends are now well understood. . . . Techniques described in this book give promise of great power in their further development. It seems not unreasonable to anticipate thoroughly literate translations of literary works as good as published run-of-the mill translations. . . . [Translation of poetry] seems hardly more extravagant now than an automatic dictionary did ten years ago. (Booth and Locke, 1955, p. 14)

Not only may machine translation of poetry be possible, it may turn out to be the best option for minority languages, the british engineer Richard Richens added: "if a Georgian speaker wishes to appreciate the imagery of Welsh poetry, machine translation might well prove to be the ideal approach" (Richens, 1958, p. 144).

Hopes were also expressed on the other side of the iron carpet. Yuri Panov from the USSR Academy of Sciences had carried out some experiments with machine translation in 1955, which he reported in a book published in 1956. The conclusion (in the 1960 English translation) reads in its entirety:

> It is, of course, a long way from the first attempts at automatic translation which have now been made in the U.S.S.R. and abroad to the practical realisation of the aim of translating by machine on any large scale. We feel, however, that the position is not as black as it is painted by some scientists abroad, and that there is every ground for expecting real successes in the immediate future, at least in the translation of scientific and technical material. (1960, p. 65)

Unlike Locke and Booth, Panov predicted that we might have to wait for some time for machine translation of literary works (1960, p. 7). Translation of scientific material, on the other hand,

> . . . is a task that we can undoubtedly hope to master with the help of automatic translation. The first steps in this direction have already been taken an real progress in this field can be looked for in the next years. (1960, p. 8)

Panov's optimism was informed by his own experience, but also by the records he had read of IBM's first experiments with machine translation. In cooperation with Georgetown University in Washington, DC, IBM had developed a Russian-to-English machine translation programme implemented on an IBM 701 computer. The machine was shown in action, translating a number of short sentences, at a public demonstration in January 7th, 1954. The IBM press release (IBM, 1954) deemed the event an "unqualified success":

> Russian was translated into English by an electronic "brain" today for the first time. . . .
> A girl who didn't understand a word of the language of the Soviets punched out the

> Russian messages on IBM cards. The "brain" dashed off its English translations on an automatic printer at the breakneck speed of two and a half lines per second.
>
> "Mi pyeryedayem mislyi posryedstvom ryechyi," the girl punched. And the 701 responded: "We transmit thoughts by means of speech." (IBM, 1954, unpaginated)

The glossy event created a fair amount of public attention. The poetic translation sample quoted in the press release made its way into many newsmedia, among others TIME Magazine, that ended their story with the following sentences:

> The computer is far from ready to translate a book from Russian to English. But, says Georgetown Scholar Leon Dostert: "Five, perhaps three, years hence, interlingual meaning conversion by electronic process ... may well be an accomplished fact." (Anonymous, 1954)

The techniques behind this "successful culmination" could in principle be extended to any language, the IBM programmer behind the implementation predicted:

> Though Russian into English translation was used as the model in the first practical experiment, the Georgetown language conversion method can be applied almost universally. (Sheridan, 1955, p. 9)

Apparently, we were even closer to realising machine translation in the early 1950s than we are now. In the light of these quotes, it is natural to ask oneself if our hopes are more justified this time around. This question is best answered by taking a closer look into the historical trajectory of machine translation from its beginning to this day. In chapter 2, I will accordingly go through the most important trends machine translation has lived through in the course of the last 60 years.

1.4 The quality of machine translation

Before I proceed, however, I would like to spend some pages examining the actual quality of present-day machine translation. I will do this by looking at actual translations as well as systematic evaluation results. This will qualify what I mean by "high quality" and "errors."

1.4.1 Some examples of the current state of the art

This subsection contains some sample machine translations. You will thus find me simply pointing and showing without any theoretical considerations. The machine translations exhibited here will give a general picture of how far the field has currently advanced, and they further provide some preliminary illustration of the theoretical problems I discuss in chapter 3.

There are several free machine translation tools available on the internet, including Babelfish (`babelfish.yahoo.com`), Google Translate (`translate.google.com`), PROMT (`www.online-translator.com/`), SDL's Free Translation tool (`www.freetranslation.com/`), Systran (`www.systran.co.uk/`), and WordLingo (`www.worldlingo.com/`).

A fairly recent study compared all of these using two different evaluation scores and two different test corpora (Kit and Wong, 2008). For the language pairs I consider in this section, Google Translate and Babelfish were the only ones consistently among the top 5% (Kit and Wong, 2008, pp. 317–18). The most recent comparison by the US National Institute of Standards and Technology rated Google Translate as the best available system with respect to three language pairs out of four tested (NIST, 2008). All of the following examples have been produced with Babelfish and Google Translate on March 25, 2009.

Both Babelfish and Google Translate give a readable output for close language pairs, but make errors in every single sentence. Take this Spanish text:

> Los años de servicio del presidente Obama responden a una creencia inquebrantable de que es posible unir a la gente alrededor de una política de propósito. En el Senado de Illinois, el presidente Obama promovió la reforma ética más importante en veinticinco años, recortó los impuestos de las familias en la fuerza laboral y amplió los servicios de salud para los niños y sus padres. Como senador por el estado de Illinois, trabajó con la oposición para moderar la influencia de los grupos de presión, asegurar las armas de destrucción masiva y promover la transparencia gubernamental reportando todos los gastos del Gobierno en Internet.

The text is grammatical, faithful to its genre, and contains solely terms that have appeared repeatedly in Spanish-language newspapers. The English translation by Babelfish goes as follows:

> The years on watch of president Obama respond to an unshakeable belief that it is possible to unite to people around an intention policy. In the Senate of Illinois, president Obama promoted the more important ethical reform in twenty-five years, he trimmed the taxes of the families in the labor force and extended the services of health for the children and their parents. Like senator by the state of Illinois, it worked with the opposition to moderate the influence of the pressure groups, to assure the arms massive destruction and to promote the governmental transparency reporting all the expenses of the Government in Internet.

There are quite a lot of errors, including some that may be potentially misleading (*on watch*, *like senator*, *assure the arms massive destruction*). Most, but not all, are due problems with prepositions (*de servico*, *como senador*, *por el estado*, *en Internet*) and pronouns (*it worked* for *trabajó*).

The Google Translate version reads:

> Service years of President Obama respond to an unwavering belief that it is possible to unite people around a politics of purpose. In the Illinois Senate, President Obama promoted major ethics reform in twenty-five years, cut taxes for families in the labor force and expanded health services for children and their parents. As a state senator from Illinois, worked with the opposition to moderate the influence of lobbyists, securing weapons of mass destruction and to promote government transparency by reporting all government spending on the Internet.

This is reasonably good, but again, there are bizarre errors in every single sentence. Articles have disappeared (*los años*, *la reforma*), sentence subjects have been lost (Who *worked with the opposition*?), bad phrases have been picked (*major* instead of *the most important*, *state senator from Illinois* instead of *senator of the state of Illinois*). The Spanish listing *moderar*, *asegurar*, *promover* has also been inconsistently translated as *to moderate*, *securing*, *to promote*. That destroys the rhythm and it makes it sound as if the moderation of lobbyist influence somehow caused the securing of weapons.

If we look at just more distant languages, the situation worsens. Consider the Dutch text:[4]

[4] Google Translate did not have the Dutch-to-English language pair at the time Kit and Wong (2008) compared the translation engines. They found Babelfish to be the best Dutch-to-English translator on both test corpora and by both measures (pp. 317–18).

> Wanneer het uiteinde van het zaagblad een voorwerp raakt of wanneer het hout naar beneden drukt en de zaagketting in de zaagsnede vastknijpt, dan kan de motorzaag een terugslag krijgen. Bij aanraking van het zaagbladuiteinde kan het zaagblad bliksemsnel worden omhoog geworpen in de richting van de gebruiker. Door het afknijpen van de zaagketting aan de bovenkant van het zaagblad kan het zaagblad razendsnel in de richting van de gebruiker worden gedrukt. In beide gevallen kan de terugslag u de controle over de motorzaag doen verliezen, wat ernstig lichamelijk letsel tot gevolg kan hebben.

Babelfish renders this:

> When the end of the saw booklet touches an object or when it wood presses down and the saw collar in the saw cut vastknijpt, then the engine saw can get a repercussion. At contact of the saw booklet end the saw booklet can be thrown fulgurating omhoog in the direction of the user. By pinching off the saw collar to the upper part of the saw booklet the saw booklet can be pressed super-fast in the direction of the user. In both cases can do the repercussion you the control concerning the engine saw losses, what can seriously physical lesion.

This is not just clumsy writing, but virtually unintelligible. Words from all word classes have been assigned absurd translations, some words are left untranslated (*vastknijpt*, *omhoog*), and none of the sentences are grammatical English. Google Translate does not fare better:

> When the tip of the blade hits an obstacle or when the wood up and down the saw chain in the cut down pinches, the chainsaw get a return. Contact between the blade tip, the blade lightning quickly thrown up in the direction of the user. By afknijpen of the saw chain to the top of the blade, the blade quickly in the direction of the user are printed. In both cases, the return the control of the chainsaw to lose some serious bodily injury can result.

I take it to be quite obvious that it would be problematic if this were an actual chain saw manual. It seems unlikely that any qualified human translator would produce mistakes of this kind.

Judging from this small set of examples, it appears that (1) machine translated texts may help deciphering a text you understand absolutely nothing of, but (2) they are burdened by frequent and numerous errors, and (3) some of these are positively harmful to the meaning of the source text. In the next section, I will try to justify this tentative assessment with references to some larger studies of the quality of machine translation.

1.4.2 Some generalisations about the current state of the art

Machine translation is hard to evaluate for the largely same reasons that it is hard to do. Automatic evaluation algorithms like the BLEU score (Papineni et al., 2002) exist, but they are built according to the same principles as the systems they evaluate and consequently share many of their flaws.

Another obstacle to evaluation is secrecy. Commercial or military organisations are not always keen on sharing their information. John Hutchins and Harry Somers write in their *Introduction to Machine Translation*:

> Most evaluations take place under contract and often under confidentiality agreements. Consequently there is little criticism of methodology (Hutchins and Somers, 1992, p. 161)

This is perhaps amplified by the competitiveness of the research community, which gives researchers an incentive to focus on achievements and downplay problems. “Uninteresting” errors might not find their way to publications, and discussions might be based on simplistic toy examples like *I like to eat* and *John usually goes home* (Dorr, 1993, ch. 8).

Consequently, it is not as easy to assess the state of the art as one might think. Most evaluations reported in the academic literature are either done by automatic means or behind the scenes. There are exceptions, however. In the following, I will review a couple of fairly recent evaluations in which the machine translated text was actually read by human beings.

The engineer D. Verne Morland from the National Cash Register Corporation (NRC) did a study in 2002, in which he asked NCR employees across the world to give feedback on a machine translated newsletter. He found that people with a poor understanding of the source language (English) generally found the translations “OK” or “practical” (Morland, 2002, section 6, unpaginated).

On the other hand, people who knew enough English to get a decent understanding from the original text were strongly sceptical, and almost a third of the subjects would have given up on the machine translation and signed up for the English original if they had not been research subjects (Morland, 2002, Appendix A, table 2). Comments like “I surely prefer to read English than bad Spanish” and “This poor German hurts in my eyes” were not uncommon (sections 3 and 6).

Another study, funded by the US Department of Defense, tested the intelligibility of the output of an English-to-Arabic machine translation system (Jones et al., 2005). The researchers gave a set of English-speaking subjects an English translation of a standard comprehension test used to test language skills. However, instead of foreign language materials, they were tested on a machine translation of an Arabic text. A control group received an ordinary translation of the same text.

The readability of the machine translation was notably worse than the standard translation: On the three different levels of difficulty that were measured, the percentage of correct answers dropped from 95% to 75%, from 91% to 76%, and from 79% to 51%, respectively. Further, “[r]eaction times are generally slower when reading MT output in comparison with human translations for each level” (Jones et al., 2005, p. 1011).

In 2007, a group consisting of mostly the same members did a new evaluation using the same methodology. The outcome was largely the same, showing an overall drop from 95% comprehension for the ordinary translation to 74% for the machine translation (Jones et al., 2007, p. 78). This time, the authors also suggested that it might be a good idea to lower the bar and not expect good translations on the highest level of difficulty:

> We have long felt that Level 2 is the natural and successful level for machine translation. The ability to present concrete factual information that can be retrieved by the reader, without requirements for understanding the style, tone, or organizational pattern used by the writer seemed to be present in the previous work. (Jones et al., 2007, p. 80)

However, “style, tone, or organizational pattern” apparently includes 26% of the factual content of the documents. And this does not include “hard” texts:

> It is worth pointing out that though we have many Level 1 questions, we are still not really testing Level 1 because the test does not contain true Level 1 documents. In future tests we wish to include Level 1 documents and questions. (Jones et al., 2007, p. 80)

So to sum up, the pattern that seems to emerge is, once again, that machine translations are less readable and less accurate than ordinary translations. They might still be useful in some situations, in particular if they are the only translation available, but they do not compare to ordinary human translation.

Chapter 2

A history of the limits of machine translation

In the introduction, I showed that machine translation currently produces relatively bad output compared to ordinary translation. In chapter 3, I will explain why we should not expect it to get much better in the future. This is a philosophical argument, but it has concrete, linguistic consequences that I give examples of in chapter 4.

In the present chapter, however, I will look back at some specific machine translation techniques that have been applied in the past. The presentation is historical and chronological, and it will explain how we arrived at the techniques that currently form the basis of machine translation. I begin where I left off in section 1.3.2, with the machine translation systems of the early 1950s.

2.1 The dictionary method

Warren weaver had already in 1949 toyed with the idea of doing machine translation one n-gram at a time, that is, by looking at series of words rather than isolated words (Weaver, 1955, p. 21). This was not very far from the method used in the Georgetown–IBM programme I referred to in section 1.3.2 (p. 11). In the following subsection, I will describe the algorithm behind that system as well as a some extensions of it that were typical at the time.

2.1.1 Window translation

The simplest possible algorithm for translating a sentence consists in looking up a likely translation of all the words, one by one. This works well in some cases such as these Dutch sentences:

(2.1) De massa van de maan is 1/81 van die van de aarde.
The mass of the moon is 1/18 of that of the earth.

(2.2) Dit boek is verkrijgbaar in winkels overal in Europa.
This book is available in shops everywhere in Europe.

Dictionary key	English 1	English 2	Code 1	Code 2	Code 2
vyelyichyina	magnitude	–	–	–	–
ugl-	coal	angle	121	–	25
-a	of	–	131	222	25
opryedyelyayetsya	is determined	–	–	–	–
otnoshyenyi-	relation	the relation	151	–	–
-yem	by	–	131	–	–
dlyin-	length	–	–	–	–
-i	of	–	131	–	25
dug-	arc	–	–	–	–
-i	of	–	131	–	25
k	to	for	121	–	23
radyius-	radius	–	–	221	–
-u	to	–	131	–	–

Table 2.1: The analysis of the Russian sentence *Vyelyichyina ugla opryedyelyayetsya otnoshyenyiyem dlyini dugi k radyiusu* (translated in the IBM-Georgetown experiment as *Magnitude of angle is determined by the relation of length of arc to radius*).

Many things can and do go wrong with such an algorithm, and it has never seriously been proposed for translation. Wrong word senses are often picked, and the word order is likely to come out wrong, especially in qualitatively different languages:

(2.3) Het sterrenbeeld is zichtbaar van juni tot november.
The constellation is visible *of June to November.

(2.4) Vandaag heb ik een boek besteld.
Today *have I a book ordered.

In the Georgetown–IBM method, these problems were addressed by, in effect, translating triples of words instead of words. The algorithm consisted of two main parts, a "lexical" part which split the Russian text up into morphemes, and an "operational" part which ran through the morphemes, picked their translations, and rearranged them (MacDonald, 1954; Sheridan, 1955).

For instance, during the showcase, the programme was fed the sentence

(2.5) Vyelyichyina ugla opryedyelyayetsya otnoshyenyiyem dlyini dugi k radyiusu.

The "lexical syntax program" split this sentence into the morphemes shown in table 2.1, using its 250-word dictionary. Each morpheme had one or two meanings and was equipped with some markup (codes 1, 2, and 3). The markup indicated the possible functions and meanings of the morpheme.

The "operational syntax" programme examined these morphemes one by one while keeping the entire previous and next word in the working memory. One of six different macros was then executed, depending on the markup attached to the three words in the memory (cf. table 2.2, page 18).

1. If the current morpheme is marked ‘110’:
   (a) If any morpheme in the preceding word is marked ‘21’:
       Use meaning 1; reverse order.
   (b) Else:
       Use meaning 1; retain order.
2. If the current morpheme is marked ‘121’:
   (a) If any morpheme in the following word is marked ‘221’:
       Use meaning 1; retain order.
   (b) If any morpheme in the following word is marked ‘222’:
       Use meaning 2; retain order.
3. If the current morpheme is marked ‘131’:
   (a) If any morpheme in the preceding word is marked ‘23’:
       Use meaning 2; retain order.
   (b) Else:
       Use meaning 1; reverse order.
4. If the current morpheme is marked ‘141’:
   (a) If any morpheme in the preceding word is marked ‘241’:
       Use meaning 1; retain order.
   (b) If any morpheme in the preceding word is marked ‘242’:
       Use meaning 2; retain order.
5. If the current morpheme is marked ‘151’:
   (a) If any morpheme in the following word is marked ‘25’:
       Use meaning 2; retain order.
   (b) Else:
       Use meaning 2; retain order.
6. If the current morpheme does not have a mark for code 1:
   Use meaning 1; retain order.

Table 2.2: The six rules governing the Georgetown–IBM system, based on the text of Sheridan (1955).

In the case of the sentence (2.5), the rules 6, 2(b), 3(b), 6, 5(b), and 3(b) would apply as the algorithm scanned through the sentence. With the morpheme under examination in capital letters, the process can be visualised thus:

1. VYELYICHYINA ugl-a opryedyelyayetsya otnoshyenyi-yem dlyin-i ...
2. magnitude UGL-a opryedyelyayetsya otnoshyenyi-yem dlyin-i ...
3. magnitude angle-A opryedyelyayetsya otnoshyenyi-yem dlyin-i ...
4. magnitude of angle OPRYEDYELYAYETSYA otnoshyenyi-yem dlyin-i ...
5. magnitude of angle is-determined OTNOSHYENYI-yem dlyin-i ...
6. magnitude of angle is-determined the-relation-YEM dlyin-i ...
7. magnitude of angle is-determined by the-relation DLYIN-i ...

And so on. The algorithm picks one of the two meanings of each morpheme and then performs a swap, a deletion, or an insertion. The markup does not carry any obvious linguistic meaning, but it evidently serves to disambiguate the morphemes and get the modifiers and case affixes into their right places. Having apparently no linguistic idea, the algorithm would have been very hard to debug or extend in any systematic way.

It would seem that the algorithm works because input and the rules were tailored for each other. The magnitude of angles and quality of coal is translates correctly, but a magnitude of coal would not be. Further, the algorithm would not be able to handle interdependent words if they were separated by adverbs, adjectives, or relative clauses.

For instance, we might write a swapping-rule like 1(a) or 3(b) to deal with French expressions like

(2.6) équations différentielles

This could reverse the order of the translated words, producing the correct translation, *differential equations*. But since the swapping may only affect the immediately preceding word, the rule would fail for expressions like

(2.7) équations différentielles linéaires

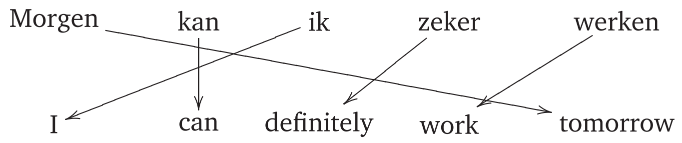
(2.8) sanctions juridiques internationales

These would turn out as *differential linear equations* and *legal international sanctions* instead of *linear differential equations* and *international legal sanctions*.

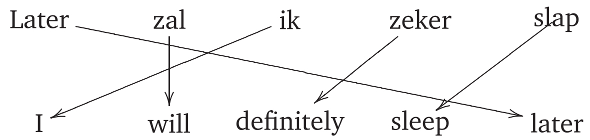
Not long after the IBM–Georgetown experiment, Andrew Booth and two of his co-workers suggested a technique that could fix this problem (Booth et al., 1958, p. 87–90). In a sense, their technique consisted in expanding the window that the IBM–Georgetown algorithm had read its input through.

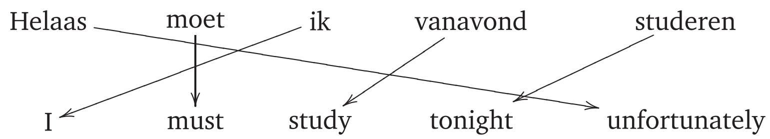
Since the rules of the Georgetown–IBM algorithm involved the words to the right and left of a given morpheme, it could be seen as an application of a translation template with three slots able to hold one word each. Since the words of the input text were divided into different classes by their markup, different templates could be used for different trigrams of word classes.

Booth et al. suggested that this idea could be generalised to translation tempates of, say, five words instead of three (Booth et al., 1958, p. 90). A particular series of word classes would then trigger a particular alignment template, as in

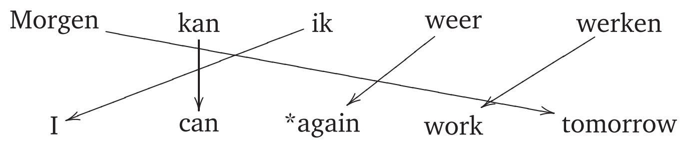
(2.9) 

Whenever similar series of words were encountered, this alignment template could be reused:

(2.10) 

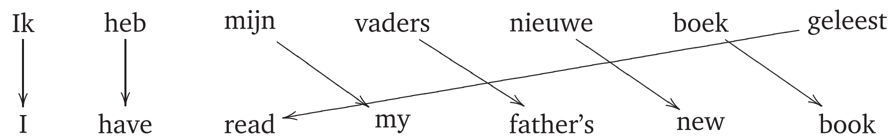
(2.11) 

However, the "similarity" of two different strings of words would have to be manually assessed. If the markup did not catch all relevant distinctions, ill-formed translations like the following might appear:

(2.12) 

The only way such an examples could be excluded was by introducing a distinction *zeker* and *weer*. Whenever more relevant distinctions were discovered, the set of markup codes would thus have to be updated or refined. There was no obvious way to test whether such a markup scheme was adequate, given that it could be.

Booth et al.'s suggested thus had little to say on when to move a word during translation. There issue of how to move the words, however, was fairly straightforward. For large enough alignment templates, it would be able to perform reorderings such as

(2.13) 

Visualised in the same way as the Georgetown–IBM example above, this translation would run through the following steps:

1. IK heb mijn vaders nieuwe boek geleest. (Retain word order.)
2. I HEB mijn vaders nieuwe boek geleest. (Retain word order.)
3. I have MIJN vaders nieuwe boek geleest. (Retain word order.)

4. I have my VADERS nieuwe boek geleest. (Retain word order.)
5. I have my father's NIEUWE boek geleest. (Retain word order.)
6. I have my father's new BOEK geleest. (Retain word order.)
7. I have my father's new book GELEEST. (Move the target expression five places left.)
8. I have read my father's new book. (Halt.)

This procedure can be seen as a translocation of the word *read* from right to left, but it can also be seen as a translocation of the phrase *my father's new book* from left to right.

If there were more verbs clustered at the end of the sentence, they could be moved left, too. A five-place alignment template would thus also allow us to translate a sentence like *I must have been trying to read my father's new book*. There is no limit on how far right a group of words can be moved using this approach. However, in the case of a five-place template, no more than four words can be moved at once.

Booth et al.'s reordering technique would is problematic for a number of reasons. For instance, *my father's new book* might have had a number of modifying clauses attached to it:

(2.14) my father's new, highly controversial book that he wrote during his time in Kenya

There is no natural boundary on the lengths of such blocks. Since the algorithm fails whenever it has to move n or more words at once using an alignment template of width n, it would fail for indefinitely many examples of this kind.

Similar problems would arise around violated incoherence constraints such as

(2.15) (a) The men *is fired.
(b) The men who build the new residence *is fired.
(c) The men who build the president's new residence *is fired.
(d) The men who build the roof on the president's new residence *is fired.
(e) The men who build ... the roof on the president's new residence *is fired.

If we read such a sentence through a window of finite width, there is no general way to discover the incoherence between *men* and *is*. This problem may be even worse in a language like Dutch, where *ze zijn* and *ze is* both may be coherent, depending on whether ze is intended to mean *they* or *she*. The context relevant to that distinction might lie any number of sentences back in a text. Apparently, then, som kind of structural analysis was called for, rather than just wider blocks.

Apart from word order and alignment, Booth et al. also hinted at ways to deal with ambiguity. In their approach, this issue too should be dealt with by refining the markup.

The lexical module should, for instance, warn the operational module if a word might mark the beginning of an idiom (Booth et al., 1958, p. 91–94). The dictionary entry for a word like *in*, for instance, should thus contain references to the expressions *in general*, *in order to*, *in the way of*, etc. Whenever the entire phrase is found, the literal translation should be overridden.

By this method, anything that looks like an idiom gets treated as one. No distinction is made between *He let down his guard* and *He let down his friends*, or between *He had blood on his hands after the trial* and *He had blood on his hands after the operation*.

That does not always do justice to the meaning of the text, but as I will argue later, it is surprisingly hard to see how else to deal with the problem. Many machine translation systems consequently accepted rather than fixed this problem. This was the case for Systran, for instance, which was written in the 1970s (Hutchins and Somers, 1992, p. 185). Even today, the publicly available versions of Systran still exhibit this insensitivity (tested August 8, 2009).

Idioms are just one species of ambiguity, though, but Booth et al. proposed that the remaining problems could be solved by even more markup (Booth et al., 1958, pp. 94–95). For instance, the French word *noyau* may be translated as *kernel*, *nut*, *centre*, *nucleus*, etc. depending on the context. To choose between these translations, the system should do a preliminary scan of the text to see whether its topic is “mathematics,” “botany,” “sociology,” or “physics” (Booth et al., 1958, p. 94):

> After some part of the text has been processed it will be found that one of the categories has been used more frequently than the others and this is taken as the indicator for deciding between the alternative meanings which occur. (Booth et al., 1958, p. 95)

An apple pit mentioned between two references to gravity would thus have to live with the translation *apple nucleus*. But the algorithm should actually be even more simplistic, Booth et al. argued:

> It is the opinion of the authors, however, that almost all ambiguities can be treated by reducing the entries to that single word which best represents the sense of all of them. (Booth et al., 1958, p. 95)

Words would thus not just be granted only one meaning per text, but only one meaning at all. Using such a method, we should expect to see translation like *The quality of coal is determined by calory content* or *Magnitude of coal is determined by length of arc to radius* (cf. MacDonald, 1954, p. 6).

2.1.2 The need for depth

While Donald Booth and his coworkers were designing translation templates, other researchers toyed with the idea of using abstract meaning representations in language processing. In a report on information retrieval, Simon M. Newman from the U.S. Patent Office wrote:

> Some aspects of research in mechanized translation of language give promise that eventually the meaning of a sentence, in context, will be mechanically extracted from the printed page, “translated” into a simple, unambiguous language, and stored in a machine memory for use in an information retrieval system. (Newman, 1957, p. 9)

In such a representation, there would be no ambiguous words, and all logical relations would be completely explicit. *The cat is on the mat* would be mapped onto a perfectly clear statement of fact—possibly of a form like $\exists x, y : cat(x) \wedge mat(y) \wedge on(x, y)$—and concepts would defined by absolute and objective criteria. This dream already had a long pedigree in Western philosophy (Maat, 2004), but with the advent of machine translation, it gained new attention and relevance.

In so far a machine translation system can translate *ugl-* as *angle* in one context and *coal* in another, it does of course use some kind of distinction between word senses. In a truly language-independent representation, however, all meanings that could possibly make a difference in any other imaginable language should be distinguished.

Such a representation has never been realised and indeed never will be for reasons I explain in chapter 3. But there were also a more practical definitions and motivations for an interlingua.

Most notably, any kind of intermediate interpretation between a set of languages would decrease the number of translation algorithms that one would need to implement.

For instance, instead of writing 5 × (5 − 1) programmes to cover all paths between English, French, German, Spanish and Russian, one could just create 2 × 5 algorithms for moving back and forth between the intermediate language and the five main languages (cf. figure 2.1).

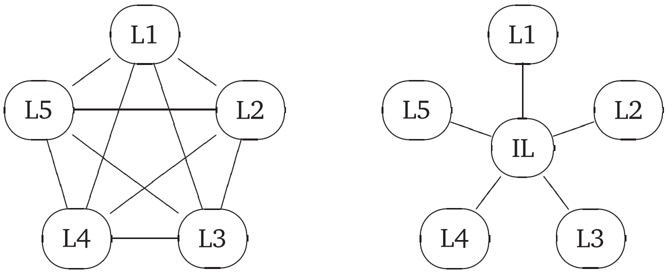


Figure 2.1: The $n \times (n-1)$ connections between a set of languages, and the $2 \times n$ connections needed when an intermediate language is used.

For that purpose, any language would do as intermediate, including one of the languages in the original set. Chinese was often suggested, and Esperanto was used in some cases (Richens and Booth, 1955, p. 25; Booth et al., 1958, p. 294).

In practice, however, things do of course not work as smoothly as such slick arguments suggest. Translation errors accumulate, and Esperanto contains just as many ambiguities and problematic features as any other language. A translation from A to B and B to C is only equivalent to a translation from A to C if none of the transitions introduce errors.

But some type of deeper analysis seemed to be necessary and lacking, already in the early 1950s. Often, good translations required radical structural change, and ambiguities were hard to deal with on the basis of just neighbouring words. As Victor Yngve noted in 1955: "If we want to translate on a sentence-for-sentence basis, we must find some method for specifying the structures of the languages" (Yngve, 1955, p. 31).

He suggested the use of Markov chains such as figure 2.2 for the description and hinted in the direction that such chains could be nested to form a more general grammar:

> If this program is carried out in its full elaboration, we are left with a number of intermediate levels of structure between the word and the sentence, such as various types of phrases and clauses (Yngve, 1955, p. 33).

However, this idea was heavily criticised by Noam Chomsky in his influential book *Syntactic Structures* (1957). He provided a different theory of linguistic structure, stated in a quasi-mathematical notation that fit well with the emerging field of language engineering. "His phrase structure and transformational notations seemed to resemble a computer program," Victor Yngve later recalled (Yngve, 2000, p. 53).

Together with some of Chomsky's later works, the book has had a huge influence on practical design matters as well as general philosophy of language. The theory is, however, fairly simple, and in the following subsection, I will run through the model it suggests.

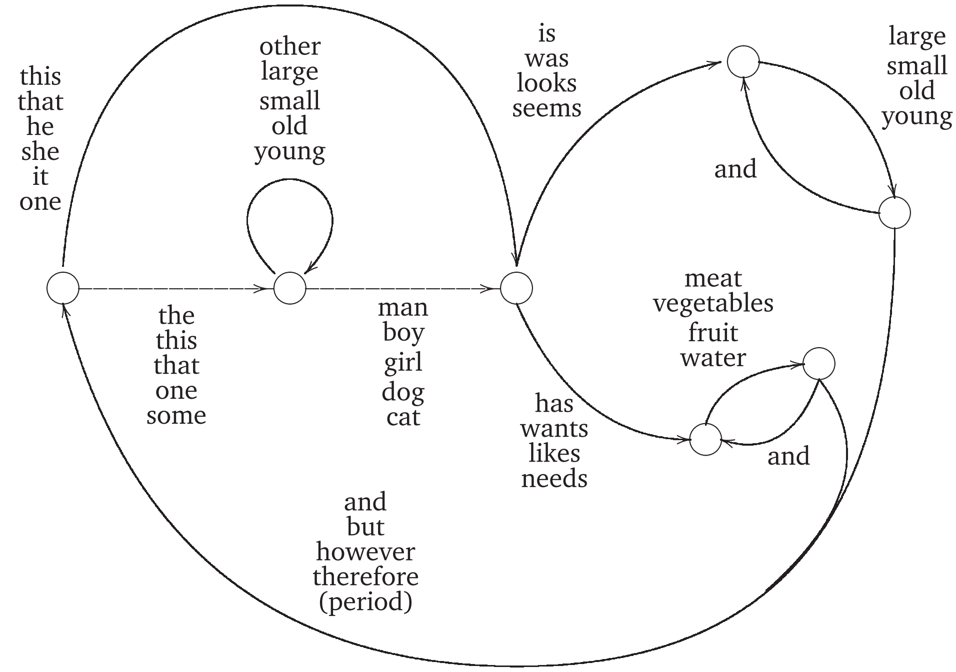


Figure 2.2: Victor Yngve's sketch of how a finite-state machine might read and produce sentences (Yngve, 1955, p. 33). By starting at the node on the left and following the arrows, one may produce sentences like *this other cat needs fruit and fruit however one is large therefore some old boy was old and large therefore. . .*

2.1.3 Phrase structure and transformational grammar

In the 1940s, the American linguist Zellig Harris had pointed out that complex constructions like *the quite old man* can be seen as expansions of simpler ones, like *the man* (Harris, 1946). Such compounds seem to fit into similar contexts or "frames" like *I saw* ________ *sleep*, and this gave Harris the idea of starting to compile rules describing all permitted substitutions. The rule $N = AN$ would, for instance, allow us to switch back and forth between *boy* and *good boy* in any sentence (Harris, 1946, §4.2).

Noam Chomsky picked up on this programme, but wanted the ambition to be more radical—he wanted to construct a set of rules that could potentially analyse all English sentences as derivates of a single unit named *Sentence*. Unlike Harris, he did not consider the terms like *A* and *N* as empty variable names, but as actual mental objects used in the process of uttering a sentence.

This had the consequence that he was working in the opposite direction of Harris, as it were: Instead of starting with actual language and heading for a set of generalisations, Chomsky started out from sets of rules, looked at the language it defined, and then tried to adjust the grammar to make the language look more like real English.

All of this becomes quite straightforward when we look at his examples. The first grammar that

Chomsky presented was the following “simple example of the new form for grammars associated with constituent analysis” (Chomsky, 1957, §4.1, p. 26):

(2.16) (a) *Sentence* $\rightarrow$ *NP* + *VP*

(b) *NP* $\rightarrow$ *T* + *N*

(c) *VP* $\rightarrow$ *Verb* + *NP*

(d) *T* $\rightarrow$ *the*

(e) *N* $\rightarrow$ *man*, *ball*, etc.

(f) *Verb* $\rightarrow$ *hit*, *took*, etc.

This set of rules was intended to be read as a “phrase structure grammar,” that is, a description of how the object *Sentence* would get elaborated into actual sentences. We should, Chomsky explained, “interpret each rule $X \rightarrow Y \ldots$ as the instruction ‘rewrite X as Y’.”

We can, for instance, derive the sentence *the man hit the ball* through the following stepwise expansion, using one rewriting rule at a time:

(2.17) (a) *Sentence*

(b) *NP* + *VP*

(c) *T* + *N* + *VP*

(d) *T* + *N* + *Verb* + *NP*

(e) *the* + *N* + *Verb* + *NP*

(f) *the* + *man* + *Verb* + *NP*

(g) *the* + *man* + *hit* + *NP*

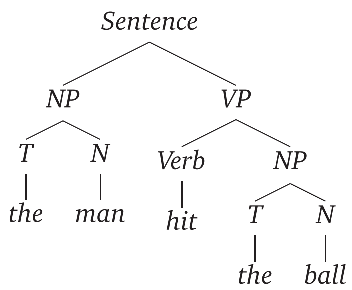
(h) *the* + *man* + *hit* + *T* + *N*

(i) *the* + *man* + *hit* + *the* + *N*

(j) *the* + *man* + *hit* + *the* + *ball*

The derived sentence can then represented as a tree showing its phrase structure:

(2.18)



Using such phrase structure grammars, Chomsky was able to draw a sharp line between grammatical and ungrammatical sentences in fairly plausible way.

However, he still had not solved the coherence problems I mentioned on page 21. The individual leaves on his phrase structure trees had no way to exchange information and could consequently not exclude sentences like *The men *hits the ball* (cf. Chomsky, 1957, pp. 17 and 76). Another quirk of phrase structure grammars was that they did not allow for the production of questions, passives, and many other sentence forms.

In order to patch these shortcomings, Chomsky created another type of grammar in which the derived sentence was only seen as an intermediate product and still had to undergo some changes before it was uttered. Passive sentences (*The ball was taken by the man*) were then explained as transformed versions of their active counterparts (*The man took the ball*).

This "transformational grammar" thus conceived of the process as a series of independent steps as in figure 2.3. In order to create such a grammar, Chomsky had to introduce new objects and write new phrase structure rules.

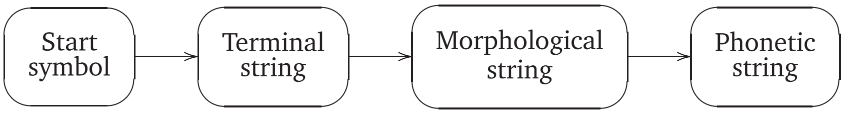


Figure 2.3: The process underlying an utterance in Chomsky's transformational model: Derivation, transformation, and morphophonemic elaboration (based on Chomsky, 1957, p. 46).

The new objects were auxiliary verbs and a set of tense, number, and person markers. The markers would not give rise to any words in the final sentence but were assumed in order to transport grammatical information. They were named *S*, *PAST*, *EN*, *ING*, and Ø (Chomsky, 1957, p. 32 and 39; I have capitalised the names to distinguish them from words).

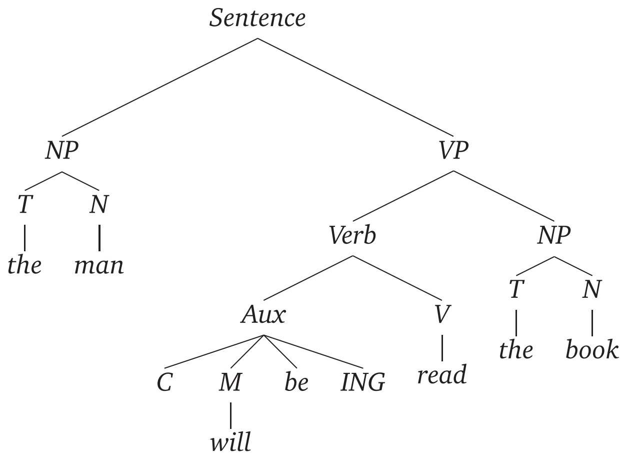
These markers were introduced into an updated phrase structure grammar (Chomsky, 1957, p. 39). In condensed form, this grammar can be written

(2.19) *Sentence* → *NP* + *C* + *(M)* + *(have* + *EN)* + *(be* + *ING)* + *V* + *NP*

where *C* is a dummy object later to be replaced by a markers (*S*, Ø, or *PAST*), *M* is a modal verb (*will*, *can*, *may*, *shall*, *must*), and *V* is a transitive verb. The elements in parentheses are optional.

To arrive at *What will the man be reading?* from this new type of grammar, we should go through the following steps: First, we derive the "kernel sentence"

(2.20)



For convenience, we can represent this tree one-dimensionally:

(2.21) *the + man + C + will + be + ING + read + the + book*

Then we begin to apply the transformations.

The first transformation turns the dummy object *C* into a verb affix congruent with the *NP* on its left, in this case *the + man* (cf. pp. 39–40). Since *the + man* is third person and singular, *C* may either be transformed into *S* or *PAST*. Given that we want the sentence to come out as present tense, we should transform it into *S* (the parentheses are meaningless and only included for readability):

(2.22) *(the + man + C) + will + be + ING + read + the + book*

(2.23) *(the + man + S) + will + be + ING + read + the + book*

Then, the sentence is turned into a question by an optional transformation named T_q (p. 63). This moves the leftmost affix–verb group (*S + will*) to the place left of the leftmost *NP* (*the + man*):

(2.24) *(the + man) + (S + will) + be + ING + read + the + book*

(2.25) *(S + will) + (the + man) + be + ing + read + the + book*

Once we have done this, we can transform the string into a *wh*-question by a two-step transformation named T_w (pp. 69–72). In the first step, we pick an *NP* and move it to the very left of the string. In the second step, we rewrite the *NP* as *who* or *what*, depending on whether it is animate.

In this case, we can either move *the + man* or *the + book*. Given the sentence we are aiming for, we should choose *the + book*:

(2.26) *S + will + the + man + be + ing + read + (the + book)*

(2.27) *(the + book) + S + will + the + man + be + ING + read*

Since *the + book* is inanimate, it is turned into *what* by the second step of T_w:

(2.28) *(the + book) + S + will + the + man + be + ING + read*

(2.29) *(what) + S + will + the + man + be + ING + read*

The last transformation to move anything is the so-called "auxiliary transformation" (p. 39). It moves any affix located left of a verb one place to the right. It also inserts a word break sign, written #:

(2.30) *what + (S + will) + the + man + be + ING + read*

(2.31) *what + (will + S) # the + man + be + ING + read*

(2.32) *what + will + S # the + man + be + (ING + read)*

(2.33) *what + will + S # the + man + be + (read + ING) #*

Before this string is turned into an actual utterance, we apply the "word boundary transformation" (p. 39). This inserts word breaks between all terms, except between verbs and affixes:

(2.34) *what + will + S # the + man + be + read + ING #*

(2.35) *what # will + S # the # man # be # read + ING #*

At this point, we have run through every applicable obligatory transformation. We can then use a set of "morphophonemic rules" to look up the sounds corresponding to the terms (Chomsky, 1957, p. 32). This produces a sound string corresponding to the sentence *what will the man be reading*.

2.1.4 The utility of syntax

Chomsky's way of thinking about linguistic structure has had a tremendous impact on the history of linguistics. Even though many of his ideas proved to be controversial, constituent analysis soon became a natural part of the vocabulary of linguistics. A decade after the publication of *Syntactic Structures*, Chomsky's research would appear as standard textbook material in publications like David Hays' *Introduction to Computational Linguistics* (Hays, 1967).[1] This reflected a shift in the perspective machine translation researchers had on their own field.

Thus, in 1968, the researcher Bernard Vauquois categorised diverse translation algorithms in terms of how much they analysed their input before generating their output (Vauquois, 1968; Hutchins and Somers, 1992, pp. 128–29; cf. also figure 2.4). A system like the Georgetown–IBM system would be seen as bottom-level, performing as little analysis of the text as necessary for translation. All the substantial work was done by the swapping, insertion, and deletion macros.

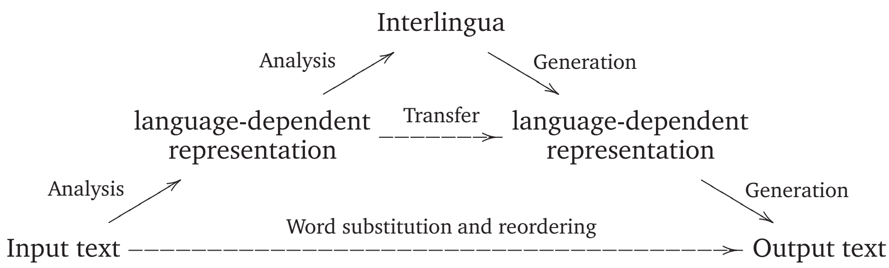


Figure 2.4: The so-called "Vauquois triangle" used for categorising machine translation systems. The figure was originally published by Vauquois (1968), but this version is based loosely on Hutchins and Somers (1992, p. 128).

Systems which were placed in the middle relied more onanalysis and less on manipulation. They might, for instance, parse an input sentence to create a phrase structure tree, move the branches around, and then translate the words at the leaves. A system at the top level would be one that first analysed its input into a language-independent representation, and then generated its output on the basis of this representation.

Vauquois' analysis of the situation caught on and is still often referred to (cf. for instance Hutchins and Somers, 1992; Kit and Wong, 2008). His discrete allusion to divine purification—placing "deep" analysis at the "top"—also foreshadowed the bias towards increasingly analytical systems that would dominate the next 20 years of research in machine translation.

In spite of this general change of philosophy, syntactical analysis did not fulfil all the hopes that machine translation researchers initially invested in it. This has several reasons.

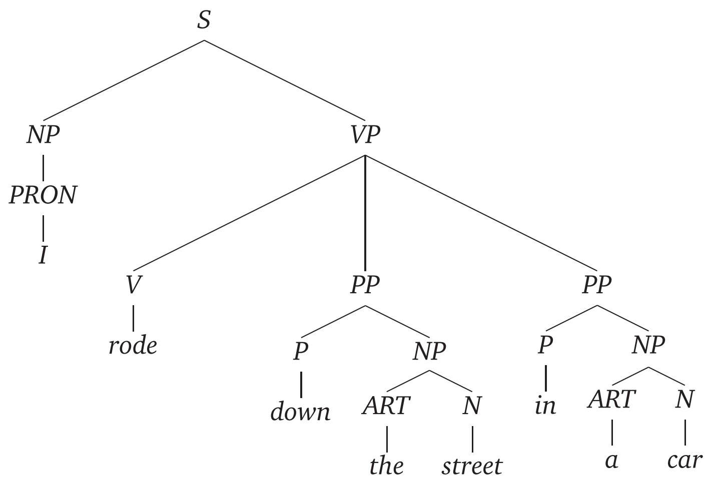
For one thing, parsing a sentence to analyse it in terms of its constituents is not as straightforward as examples like *the man hit the ball* suggest; almost any sentence of average length is subject to some structural ambiguity. Take for instance the following sentence (Winograd, 1971, p. 89):

[1] Suprisingly, the idea of the transformation itself turned out to be one of the most controversial aspects of transformational grammar. Victor Yngve criticied it heavily for being completely computationally unfeasible; the so-called generalised phrase structure grammar replaced it by "metarules" that could generate a new rewriting rules in a systematic way; lexical-functional grammar replaced it by coherence constraints on the trees; Ronald Langacker criticised it for having misunderstood the difference in perspective between active and passive sentences.

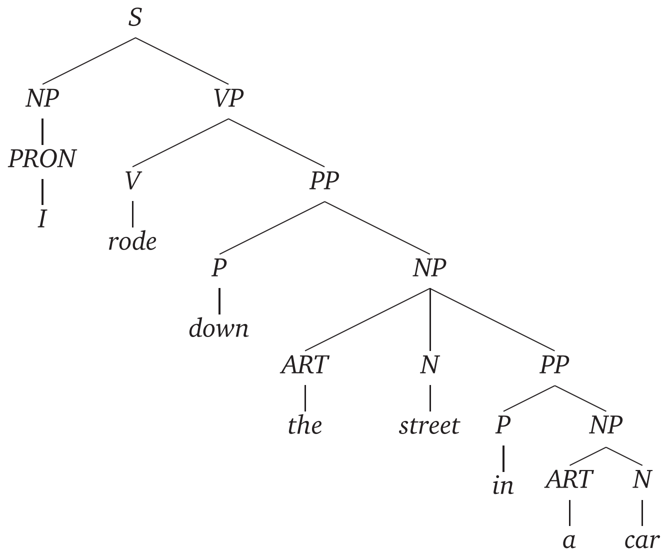
(2.36) I rode down the street in a car.

The constituent *in a car* can be attached to two different nodes in the parse tree of this sentence. Depending on the exact analytical framework, the resulting trees would look more or less as follows:

(2.37)



(2.38)



This corresponds to two radically different readings: In (2.37), *in a car* modifies the verb phrase, so it is the riding down the street that takes place in a car. In (2.38), it is *the street* which is modified by *in a car*, so that the street is *in a car* like a hand is in a pocket.

This is of course an absurd reading. But the problem is that an abundance of such readings are often available: *The woman wanted the dress for her daughter* or *Mary saw the man who had lived with her while on maternity leave* exhibit similar ambiguities with respect to the phrases *for her daughter* and *while on maternity leave*. This problem in its general form is known as the “prepositional phrase attachment problem” (cf. Kimball, 1973, §4.2; Wilks et al., 1985).

This, in its turn, is just one example of the kind of ambiguities that occur when constituents may be added to a construction from the right. Coordinations give rise to similar problems:

(2.39) (a) I noticed that he was shaking and didn't say anything. (= he didn't)

(b) I noticed that he was shaking but didn't say anything. (= I didn't)

Similar examples could be given with conjoined noun phrases instead of conjoined verb phrases. Relative clauses also often cause such problems, as in the following example (Hobbs, 1976, p. 23):

(2.40) Archeologists found small flint blades with parallel sides, struck serially from the cores, which had assumed conical shape by the time they were discarded.

The relative clause *which had assumed ... they were discarded* can be attached to several nodes in the parse tree: It may modify the *cores*, the *sides*, or the *blades*.

Phrases that attach from the left are just as problematic as phrases that attach from the right. For instance, the modifier *government* clearly has a different scope in the following sentences:

(2.41) (a) He worked with government agencies and offices. (= government offices)

(b) He worked with government agencies and NGOs. ($\neq$ government NGOs)

Similar scope problems also occur for adverbs and adjectives, as they do in

(2.42) a pretty original movie's screenplay

This construction has different analyses depending on whether *pretty* modifies *original*, *movie*, or *screenplay*, and whether *original* modifies *movie* or *screenplay*. The result is not just one or two, but five qualitatively different readings.

Apparently, then, we constantly walk around in a forest of parse trees when we hear simple sentences like *I rode down the street in a car*. Even so, we mostly seem to shave off the wrong readings automatically and instantaneously. This suggests that "pure syntax" may after all not be as "pure" as it seems.

There is also another reason why syntactical analysis could not fulfil all the ambitions of machine translation: Even when there is just one single structural reading of a sentence, there might still be a lot of ambiguity left on the lexical level.

The reason that syntax at one point seemed to promise otherwise is that there are cases in which structural analysis does exclude some word senses. For instance, *hatch* can either be a noun (meaning "an opening of restricted size") or a verb (meaning "brood, incubate"). But not all frames allow for both parts of speech:

(2.43) The hatch closed.

No derivational or transformational history can give rise to the sentence *T # Verb # Verb + PAST*. We thus have to read the sentence as *T # N # Verb + PAST*, interpreting *hatch* as a noun.

Unfortunately, this is a quite rare kind of ambiguity. If we take the meanings recorded in the semantic database WordNet seriously as representations of actual polysemy, only around 1.7% of all English words are candidates for syntactical disambiguation (cf. table 2.3). In actual text, such as the Brown corpus, this number is higher, but so is the number of polysemous words in general (cf. table 2.4; the numbers have been produced using WordNet's build-in stemming tool).

The potential utility of syntactical disambiguation is thus quite low relative to the need for other types of disambiguation. Parsing similar texts, we should expect to run into many disambiguation problems where the syntax provides no additional clues.

Word senses	In > 1 word class	In 1 word class	Total
> 1 in some word class	4,954 (3.4%)	19,474 (13.2%)	24,428 (16.6%)
0 or 1 per word class	2,445 (1.7%)	120,433 (81.8%)	122,878 (83.4%)
Total	7,399 (5.0%)	139,907 (95.0%)	147,306 (100.0%)

Table 2.3: Two different kinds of ambiguity in WordNet 3.0, between-part-of-speech ambiguity and within-part-of-speech ambiguity. The word *trickle*, for instance, has one sense as a noun and one sense as a verb, while *scripture* has two senses as a noun and none as any other part of speech.

Word senses	In > 1 word class	In 1 word class	Total
> 1 in some word class	51,312 (48.6%)	38,637 (36.6%)	89,949 (85.2%)
0 or 1 per word class	3,279 (3.1%)	12,376 (11.7%)	15,655 (14.8%)
Total	54,591 (51.7%)	51,013 (48.3%)	105,604 (100.0%)

Table 2.4: Ambiguities in the "learned" ("J") section of the Brown corpus. 105,604 of the 181,888 tokens in the section were successfully stemmed. The table shows how many possible word classes and word senses each token might represent.

2.1.5 "Slow and painful reading"

During the first years of research in machine translation, substantial amounts of money were dedicated to the new field, both by non-governmental entities such as IBM and the RAND corporation, and by institutions like the CIA and the American Department of Defense. In 1964, however, a special committee was formed in order to assess the progress of the research, the Automatic Language Processing Advisory Committee (ALPAC).

The result was an official report (Pierce et al., 1966) that surveyed the state of the art in machine translation. Its conclusions were mainly negative to the extent that it "effectively killed machine translation research in the States" (King, 1984, p. 352)

But the report did have quite a bleak view of the state of the machine translation of its day:

> Unedited machine output from scientific text is decipherable for the most part, but it is sometimes misleading and sometimes wrong (as is postedited output to a lesser extent), and it makes slow and painful reading. (Pierce et al., 1966, p. 19)

The committee acknowledged that early promise of "accurate, readable" machine translations was alluring (p. 24), but they warned that it was most likely illusory, too:

> Early machine translations of simple or selected text, such as those [used in the Georgetown-IBM experiment] were as deceptively encouraging as "machine translations" of general scientific text have been uniformly discouraging. (p. 24)

In another chapter of the report, several examples of the quality of machine output are displayed. One "typical" sample (p. 20) comes from an article on Russian space aviation, translated by four different corporations: The Bunker-Ramo Corporation; Computer Concepts Inc.; the Foreign Technology Division of the U.S. Air Force; and the Italian EURATOM, which is stated to be "essentially the Georgetown MT system."

The first two sentences of the machine translation from the U.S. Air Force reads:

> Biological experiments, conducted on different space aircraft/vehicles, astrophyscial space research and flights of soviet and American astronauts with/from sufficient convincingness showed that short-term orbital flights lower than radiation belts of earth in the absence of heightened solar activity in radiation ratio are safe. Obtained by astronauts of dose radiation at the expense of primary cosmic radiation belt are so small that cannot render harmful influence on organism of person. (p. 22)

With some background knowledge, it is possible to reconstruct some of the original meaning of the text by making informed guesses. But by and large, the prose is ambiguous, clumsy, and ungrammatical. The "sufficient convincingness" and "organism of person" are comical albeit intelligible, but the strangely placed "in radiation ratio" and the obscure "dose radiation at the expense of primary cosmic radiation belt" are seriously ambiguous and problematic. None of the other transcripts fare any better on these points.

The scepticism expressed in the ALPAC report had already had a proponent since the late 1950s, namely Yehoshua Bar-Hillel, who had worked professionally with machine translation at MIT since 1951 (Bar-Hillel, 1964d, p. 5–6). He himself had been quite optimistic about the prospects at the time, but by 1962, his attitude had reversed: "Autonomous, high-quality machine translation between natural languages according to rigid algorithms may safely be considered dead," he proclaimed (Bar-Hillel, 1964b, p. 218).

He attributed the problems to the lack of common-sense knowledge in the computers:

> A complete automation of [the process of translation] is wholly utopian, since the fact that books and papers are usually written for readers with a certain background knowledge and an ability for logical deduction and plausible reasoning cannot be over-ridden by even the cleverest utilization of all formal features of discourse. The hopes to the contrary that which many of us had a decade ago, just turned out to be by large unrealizable. (Bar-Hillel, 1964c, p. 183)

In a short 1960 paper (Bar-Hillel, 1964a), he used the sentence "the box was in the pen" (p. 175) to illustrate this deficiency and claimed that

> . . . no existing or imaginable program will enable an electronic computer to determine that the word *pen* in the given sentence within the given context has the second of the above meanings [1: a certain writing utensil; 2: an enclosure where small children can play], whereas every reader with a sufficient knowledge of English will do this "automatically." (Bar-Hillel, 1964a, p. 175)

This problem would be solved if the computer had our knowledge as well as the skills to use it. But just knowing that there are 8 chairs in a room, Bar-Hillel argued, we also know that there are more than 5, less than 12, and infinitely many other things:

> We know all these additional facts by inferences we are able to perform, at least in this particular case, instantaneously, and it is clear that they are not, in any serious sense,

> stored in our memory. Though one could envisage that a machine would be capable of performing the same inferences, there exists so far no serious proposal for a scheme that would make a machine perform such inferences in the same or similar circumstances under which an intelligent human being would perform them. (Bar-Hillel, 1964a, p. 177)

In other words, the problem is not so much that we cannot copy the verbal inference in a computer, but that human beings apparently choose which inferences to perform by means of some other, more mysterious kind of "inference." Since this competence is hard if not impossible to simulate, so is translation.

Bar-Hillel's arguments against the possibility of machine translations were both partly wrong and partly right. As we shall see in the next section, the idea of inserting some kind of encyclopaedia into the translation algorithms was not as unrealistic or "utterly chimerical" as Bar-Hillel sometimes suggested (Bar-Hillel, 1964a, p. 176). On the other hand, his observation that there is a larger problem hidden in non-verbal human reasoning skills is significant and true. This topic will receive more attention in chapter 3, in particular in section 3.3.

2.2 Knowledge-based methods

2.2.1 If the baby does not thrive on raw milk, boil it

As we saw in the previous section, resolution of ambiguities is to some extent a matter of syntactical knowledge, but other types of knowledge come into play, too. In his 1971 doctoral thesis, the MIT researcher Terry Winograd used examples like (2.36) to argue that human beings have an "ability to integrate semantics with syntax" (Winograd, 1971, p. 22). He added that "good computer handling of language will not be possible unless computers can do as well" (Winograd, 1971, p. 23).

In other words, computers will need semantic knowledge—cars have denders, buildings have roofs, doors can open, etc.—in order to do proper natural language processing. But not only that; in many cases, the semantic, conceptual knowledge needs to be complemented with quite involved empirical knowledge about the world.

In a 1976 article on pronoun resolution, the computer scientist Jerry Hobbs argued at length that disambiguation sometimes calls for both subtle and complex inferences. His examples included this subsection's title as well as the following sentence pair (1976, p. 43–44), adapted from Winograd:

(2.44) (a) The city councilmen refused the demonstrators a permit because they feared violence.

(b) The city councilmen refused the demonstrators a permit because they advocated violence.

In these sentences, *they* refer to different antecedents, but neither syntax nor dictionary definitions can explain this difference. In Hobbs' words:

> We understand this because of our sophisticated knowledge of councilmen, demonstrators, and politics—no set of syntactic or semantic rules could interpret this pronoun reference without using knowledge of the world. (Hobbs, 1976, p. 33)

Observations of this kind let Hobbs to speculate whether it might be possible to formalise semantic knowledge and inferences as a computational method.

He had already devised an algorithm that could search for the referent of a pronoun by considering all possible candidates in a certain order, based on a constituent analysis of the sentence. This worked fairly well (Hobbs, 1976, pp. 25–28), but the algorithm failed in some cases in which a human being could effortlessly succeed.

For instance, the algorithm would pick the wrong referent of *it* in the second of these sentences:

(2.45) The boy walked into the bank. Moments later he was seen on its roof.

Because of their respective locations in the parse tree, the algorithm would give a higher priority to *the boy* than to *the bank*, in effect interpreting the sentence as *Moments later the boy was seen on the boy's roof*.[2]

But a boy cannot stand on his roof for the obvious reason that he does not have one. If we can make that argument, Hobbs figured, a computer might, too, and he proceeded to outline an imaginary programme equipped with a machine-readable knowledge base and means for making inferences. For every pronoun–referent pair considered, this programme should initiate a bidirectional search for contradictions that could exclude the candidate.

In the case of the boy on the roof, such a programme would first derive a series of consequences from its knowledge of the boy as well as premises sufficient for something to have a roof:

(2.46) $boy(x_0) \Longrightarrow male\_person(x_0) \Longrightarrow person(x_0) \Longrightarrow living\_thing(x_0)$

(2.47) $\exists y : roof(y, x_0) \Longrightarrow building(x_0) \Longrightarrow construction(x_0) \Longrightarrow artefact(x_0)$

If programmed well enough, such a process should eventually find a contradiction and terminate at a string of consequences similar to this:

(2.48) $boy(x_0) \Longrightarrow male\_person(x_0) \Longrightarrow person(x_0) \Longrightarrow living\_thing(x_0) \Longrightarrow$
$\neg artefact(x_0) \Longrightarrow \neg construction(x_0) \Longrightarrow \neg building(x_0) \Longrightarrow \neg(\exists y\colon roof(y, x_0))$

Or, put in plain language, if x_0 is a boy, x_0 does not have a roof. *it* and *a boy* consequently refer to different entities. On the other hand, having a roof is perfectly consistent with being a bank, so no amount of searching will derive a contradiction from $bank(x_0)$ and $\exists y\colon roof(y, x_0)$. *the bank* would thus accepted as the referent of *it*, and the two objects would be merged into one in the formal representation.

2.2.2 From artificial intelligence to the weather forecast

Or so the story goes. But in order to actually implement a system such as the one Jerry Hobbs imagined in 1976, a database about boys, buildings, and roofs would have to be created.

In the late 1960s, both Eugene Charniak and Terry Winograd had constructed programmes equipped with such databases at MIT (Charniak, 1972; Winograd, 1971). However, the amount of knowledge stored in their systems was so small that they could not be used for the kind of reasoning Hobbs had in mind. Winograd's dialogue system SHRDLU, for instance, only contained information on 13 objects—a table, a box, the user, the system, its arm, and eight toy blocks. The blocks, in their turn, only had a few properties—a colour, a shape, a corner position coordinate, and another corner position coordinate (Winograd, 1971, §7.1 and Appendix F).

[2]For the sake of the argument, I assume here, as did Hobbs, that *the boy* can here be referred to as both *it* and *he*.

During the 1970s, other projects expanded their frame of reference to something slightly more comprehensive. At Yale University, for instance, James Meehan created a programme called TALE-SPIN which produced small texts intended to look like the stories from Aesop's fables (Meehan, 1977, 1981). When executed, the programme would pick a number of actors, like Joe the Bear or Henry the Crow, and equip them with desires that would initiate the drama:

> One day Joe Bear was hungry. He asked his friend Irving Bird where some honey was. Irving told him there was a beehive in the oak tree. Joe walked to the oak tree. He ate the beehive. (Meehan, 1977, p. 91–92)

It also produced other considerably less well-formed stories, but the success rate was high enough for the system to earn some fame.

In machine translation, one of the first attempts at such an approach was made by TAUM, the machine translation group at the University of Montreal. In the mid-1970s, they began developing a system that could translate short weather bulletins from English to French and *vice versa* (Isabelle, 1987; Tucker, 1987, §2.2.2; Hutchins and Somers, 1992, ch. 12). The simplicity of texts like example (2.49) made them a promising field for machine translation research.

(2.49) `CLOUDY WITH A CHANCE OF SHOWERS TODAY AND THURSDAY.`

```
LOW TONIGHT 4.  H1GH THURSDAY 10.
OUTLOOK FOR FRIDAY...  SUNNY.
```

Because of this formularity, TAUM could postpone the task of writing a phrase structure grammar of the entire English language. Instead, they could stick to a much smaller set of highly idiosyncratic phrases like Canadian place names, temperature spans, dates, times, and weather conditions.

In other words, by constraining their input, the TAUM group were able to stay clear of most of the ambiguity issues that would normally arise in translation. Their system could then translate example (2.49), almost appropriately, as

(2.50) `NUAGEUX AVEC POSSIBILITE D AVERSES AUJOURD HUI ET JEUDI`

```
MINIMUM CE SOIR 4.
H1GH THURSDAY 10.
APERCU POUR VENDREDI...  ENSOLEILLE.
```

Note, though, that the third sentence was left untranslated because of the misspelling of `HIGH` as `H1GH` (Hutchins and Somers, 1992, p. 209–210).

In spite of its heuristic character and minimal coverage, TAUM's translation system was generally seen as success story. As late as 1987, it was described as "the closest approximation to fully automated high quality machine translation among currently operational systems" (Tucker, 1987, p. 30). If knowing the context of a text could improve machine translation that drastically in a restricted domain, maybe more knowledge could improve the quality of machine translation in a less restricted domain.

This challenge was picked up during the 1980s by, among others, the Center for Machine Translation at Carnegie-Mellon University. Inspired by the various artificial intelligence projects at Yale University, the Center began working on knowledge-based machine translation projects that would later result in translation systems such as TRANSLATOR, KBMT-89, KANT, and PANGLOSS.

2.2.3 Dictionary knowledge and the lexical-functional filter

In these systems, the translation process was broken into a series of independent modules. The input sentences were first analysed syntactically, then in terms of valency and functional roles, and lastly in terms of real-world knowledge. This allowed for two major types of consistency checks.

The first was applied as the programme took the step from a syntactic structure to a functional structure. It filtered out parse trees that postulated incoherent clauses. For instance, if the following clauses were postulated as sentence constituents, the responsible interpretation would be rejected:

(2.51) (a) I sleeps

(b) going to slept

(c) two letter

As we saw in section 2.1.3, Chomsky dealt with such incoherences by introducing tense, number, and person markers which were transported around the sentence by the transformations. However, since the lexical-functional grammar notation had been developed at Xerox in the 1980s (Kaplan and Bresnan, 1982), consensus had slowly shifted. Sentences like (2.51 a–2.51 c) were increasingly often considered functionally ill-formed but, in a certain sense, syntactically well-formed.

Lexical-functional grammar is a type unification grammar. That means that the individual words of a functional structure (like an object or sentence) are seen as contributors of information about the structure. If all the information can be unified into a consistent pool of information, the phrase may be interpreted as a well-formed functional structure.

For instance, to check whether *two letter* is a functionally well-formed structure, a programme should look for any information about gender, number, definiteness, etc. encoded in the words *two* and *letter*. Then, it should check whether the collected information was consistent. Since *two* signals that the noun phrase is in plural, while *letter* signals singular, the gathered information cannot be unified into a functional structure. *two letter* is thus not a coherent functional structure—as opposed to *one letter* or *two letters*, in which the number markers on the individual words cohere.

For other types of structures, other attributes would be relevant. The finite verbs have to cohere with sentence subjects, a relation dependent on both number, person, and tense values. In German, Dutch, Danish, and French, adjectives are inflected differently depending on the noun they modify. This relation, too, can be coherent or incoherent.

Another family of coherence constraints is even more lexically sensitive: Some words, most notably verbs, influence their context by allowing or disallowing certain constructions:

(2.52) (a) I dissolved the sugar.

(b) I disappeared *the sugar.

(2.53) (a) She says that we're starting now.

(b) She disturbs *that we're starting now.

(2.54) (a) I'm glad to help.

(b) I'm tired *to help.

In lexical-functional grammar, the starred constructions are thought of as complements or objects which are not solicited by the words *disappear, disturb*, or *tired*. In order to capture these constraints, we need a rich and comprehensive lexicon with information on which verbs take which objects, which words allow for which kinds of complements, etc.

Once a functional structure has been assembled, there is a variety of possible ways to represent it. In the Center for Machine Translation's KBMT-89 system, nested Lisp lists were used (Nirenburg, 1989; Goodman and Nirenburg, 1991). The items on the lists were either functional substructures or, ultimately, attribute–value pairs. A sentence like *The chihuahua ate the apple* was thus represented as a list of the following form (cf. Onyshkevich and Nirenburg, 1995, p. 19–20):[3]

```
((ROOT +EAT-V1)
 (MOOD DECL)
 (VOICE ACTIVE)
 (NUMBER $3)
 (CAT V)
 (TENSE PAST)
 (FORM FINITE)
 (SUBJ ((ROOT +CHIHUAHUA-NI)
        (NUMBER $3)
        (CAT N)
        (PROPER -)
        (COUNT +)
        (CASE NOM)
        (DET ((ROOT +THE-DETI)
              (CAT DET)))
  (OBJ  ((ROOT +APPLE-NI)
        (NUMBER $3)
        (CAT N)
        (PROPER -)
        (COUNT +)
        (DET ((ROOT +A-DETI)
              (CAT DET))))))
```

The top level functional structure is here a sentence with a subject and object as immediate substructures. It further has a number of attributes, including a tense (`TENSE`), a subject person/number (`NUMBER`). We can also see that the subject is a count noun (`COUNT +`), but not a proper name (`PROPER -`). Every structure further has an attribute called `ROOT`, which contains the syntactic head of the structure—in the case of the sentence structure, the value `+EAT-V1`, one of the verb meanings of *eat*.

2.2.4 Consistency as yet another filter

The lexical-functional coherence check thus acted as filter that could reject some parses of a sentence, just like a syntactical parse could filter out some proposed word meanings (cf. section 2.1.4). However, the truly new component in the knowledge-based systems of the late 1980s was the use of meaning representations. This enabled the use of another filter that used contextual knowledge to propose or reject certain readings.

[3]The two following examples are adapted from a 1995 article on the MIKROKOSMOS analysis system (Onyshkevich and Nirenburg, 1995), not the KBMT-89 translation system. In terms of analysis and representation, though, the systems seem to be cousins, if not twins.

The meaning representation was intended to be a “language-neutral format” or “interlingua,” capturing the meaning of the sentence in a maximally abstract way (Onyshkevich and Nirenburg, 1995, p. 5). The interlingua representation did not refer to grammatical categories such as “verb,” “subject,” or “phrase” and had in principle no unresolved pronouns. Instead, it consisted of references to actions, actors, places, instruments, etc., but given in a format much like the nested functional structures. *The chihuahua ate the apple* was thus represented by the list (Onyshkevich and Nirenburg, 1995, p. 17–18)

```
(DOG323   (INSTANCE-OF *DOG)
          (SUBSPECIES "CHIHUAHUA"))

(APPLE23  (INSTANCE-OF *APPLE))

(INGEST17 (INSTANCE-OF *INGEST)
          (AGENT (VALUE DOG323)
                 (SEM *ANIMAL))
          (THEME (VALUE APPLE23)
                 (SEM *INGESTIBLE)))
```

The KBMT-89 system thus formalised assertions by postulating the necessary objects (`DOG323` and `APPLE23`). These objects were then connected in an event (`INGEST17`).

In this case, the lists show that we do not have any specific information about `APPLE23` other than the fact that it is an instance of the concept `*APPLE`. We do, however, know that `DOG323` is of the subspecies `"CHIHUAHUA"`. Any further information—such as the fact that `DOG323` has fur or cannot breathe under water—would have to be inferred from general, stored facts about the concept `*DOG`.

Like the functional structures, meaning interpretations might turn out to be inconsistent and hence unfeasible. A dog of the subspecies “chihuahua” is a reasonable entity to postulate, and the event of this chihuahua eating an apple is not unlikely, either. But in a sentence such as *I gave her a present*, it would seem—for semantic reasons—absurd to read the word *present* as meaning “the period of time now occurring.” Nothing on the syntactic or functional level would reveal that, but an attempt to construct a meaning representation might.

Consequently, consistency in meaning could be applied as yet another constraint on top of the syntactical and lexical-functional filters (Nirenburg, 1989). Accordingly, as noun phrases were encountered in a source text, entities were postulated, and as sentences were interpreted, they were equipped with attributes. The following toy discourse served as a concrete example in a 1988 article by Jamie Carbonell and Ralf Brown from Carnegie-Mellon University (Carbonell and Brown, 1988, p. 99):

(2.55) The doctor gave John a glass of water.

John drank it.

He gave him an aspirin.

As this text was processed, the software postulated a number of entities but eventually decided that *him* and the two occurrences of *John* all referred to the same (male, human, named) entity. It also postulated the existence of a doctor of unknown gender. Then, as the last line was processed, it decided that *He* had to be coreferent with the doctor, inferring that the doctor was male.

In a sense, this procedure was a practical implementation of the “knitting” process that Jerry Hobbs had imagined (Hobbs, 1976, pp. 36–37).

2.2.5 All the knowledge in the world

A consistency check could reject readings that no syntactical or even lexical-functional constraint would rule out. So far, however, I have only described this process on the logical level, with brief mention of the actual premisses that go into an inference. This is quite consistent with much of the literature of the subject, but it also belies the actual substance of the subject. By the mid-1980s, few people considered machine representation and inference a serious obstacle to knowledge-based machine translation. The hard part was compiling the relevant knowledge—writing down everything there is to know about cars, boys, buildings, and demonstrators.

This compilation work is tedious and hard to systematise. Some early artificial intelligence projects included knowledge banks of considerable sizes, but they were nowhere near completeness or comprehensibility. Eugene Charniak from the MIT, for instance, explains in his dissertation on processing of children's stories that his strategy was simply to "start out just writing down various things we know about piggy banks" (Charniak, 1972, p. 96)—hardly a methodology that would work for large-scale projects like PANGLOSS, which employed a dozen full-time researchers at three different universities (Frederking et al., 1994).

Instead, the guiding principle became the development of "ontologies"—databases with focus on the relations between concepts. The KBMT-89 project's ontology, for instance, was centred around computer manual terms and included the fact that a `*protection-sheet` is a `*protection-material` which is a `*packaging-material`, and that the complex event `*to-push-button` has the simple parts `*push-button`, `*hold-down-button`, and `*release-button` (Gates et al., 1989, pp. 73–75).

A more well-known and still widely used ontology is WordNet, which was developed by a Princeton University group under George A. Miller and presented in 1985 (Fellbaum, 1998). WordNet consists of a large list of concepts ordered in hierarchies representing semantic relations—mostly transitive, asymmetric relations. Central to the database is a hypernym/hyponym tree which contains, for instance, the fact that a "goldfish" is a "cyprinid fish" or that a "pasta sauce" is a "sauce." In addition, some words are equipped with other information, such as the fact that a "kick stand" is a part of a "motorcycle," or that "Capital" is an instance of the concept "book." A small part of such a transitive hierarchy of words is depicted in figure 2.5 on page 40.

Such facts can be used for both negative and positive points. On one hand, we can, for instance, see that *matchbox* is a hyponym of *artefact* and hence inanimate, thus disqualifying it as a plausible referent of *he* or *she* in most situations.

On the other hand, it provides constructive information that can be used for disambiguation and interpretation. For instance, only one sense of *bed* has *mattress* as a part meronym, so in the discourse below, we might be able make the thematic connection between *the bed* and *the mattress* and disambiguate them both using WordNet:

(2.56) He laid down on the bed. The mattress was soft.

The current version of WordNet (3.0) contains more than 150,000 lemmas, and many such neat example can be given. The question is, of course, if the methods apply as generally and unproblematically as these examples suggest.

2.2.6 "Does not there is no one type of difference"

The quality achieved by the wave of knowledge-based machine translation around 1990 was never very good, at least not for texts like newspaper articles. The high-profile KBMT-89 system was tested

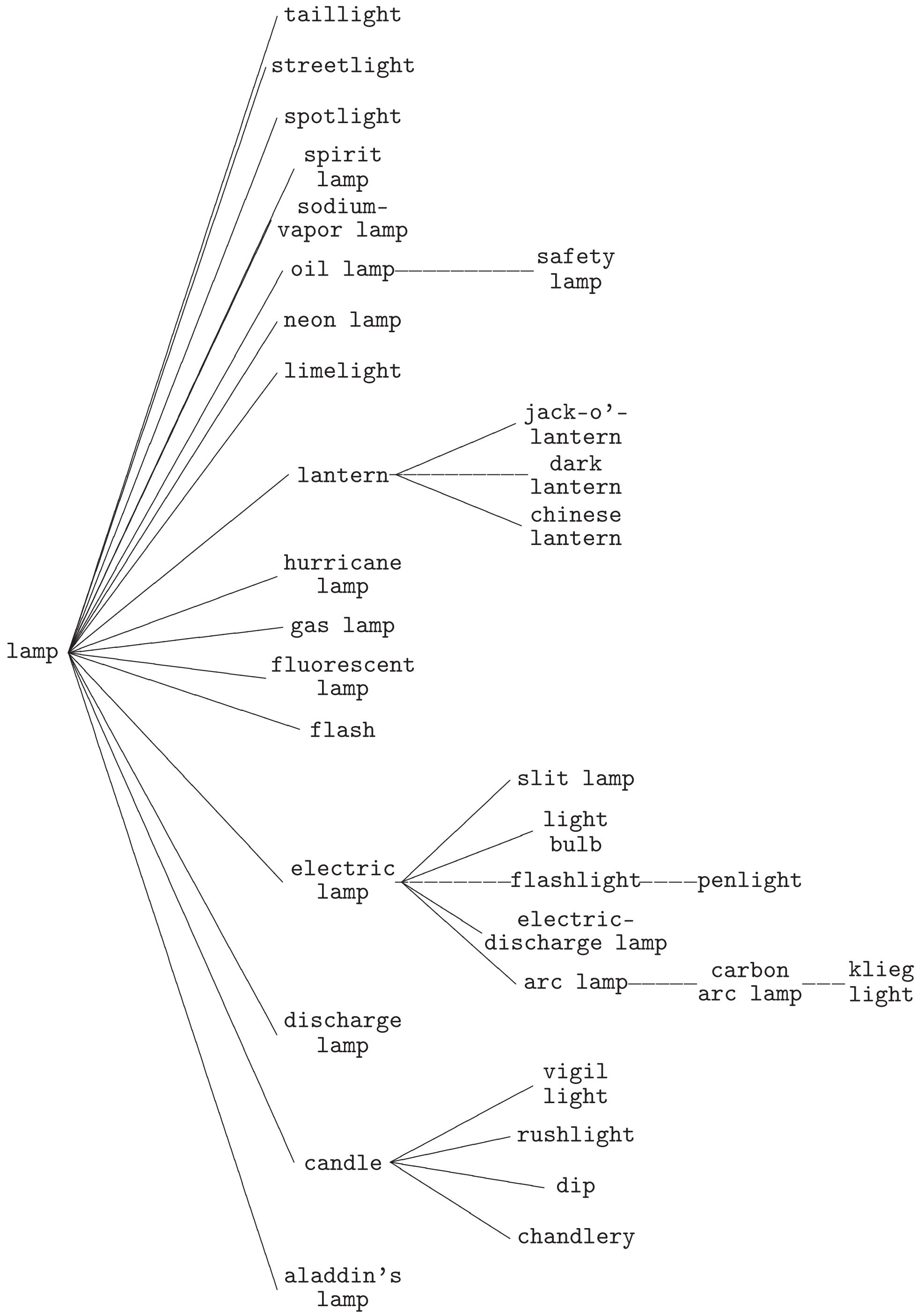


Figure 2.5: An excerpt from the bottom of the hyper/hyponym hierarchy in WordNet.

in 1991–92 and "did OK when it had lexical coverage," the research director Sergei Nirenburg told me in correspondence (July 13th, 2009). However, none of the translations produced by the system have been saved, and "the book is probably the best record remaining from that project," he stated.

The book in question is his monograph written with Kenneth Goodman (Goodman and Nirenburg, 1991), a description of the system with only two concrete sample translations, one English-to-Japanese and one Japanese-to-English (Goodman and Nirenburg, 1991, appendices A and B).

The ambitious PANGLOSS project was funded by the military research agency ARPA, who also tested it in 1992 and 1993 (White and O'Connell, 1994). According to the principal investigator, Robert Frederking, the agency focused solely on "unrestricted-domain translation, evaluating us on random newswire articles (and only translating into English)," (personal communication, October 28th, 2009). They were, for instance, given the following Spanish text:

> Menem Dijo que los Argentinos no Tienen Xenofobia contra los Paraguayos
>
> El presidente argentino, Carlos Menem, desacreditó este domingo la versión sobre la existencia en su país de algún tipo de persecución contra inmigrantes paraguayos. "De qué xenofobia me hablan", dijo el mandatario cuando fue preguntado por periodistas paraguayos por qué los argentinos "odian" a los ciudadanos inmigrantes de este país. Menem aseveró que "no hay ningún tipo de diferencias" entre argentinos y paraguayos. "Vayan a la Argentina y pregunten a cualquiera de los un millón de paraguayos residentes en mi país si son o no perseguidos y si tienen algún tipo de problemas en la República Argentina", dijo Menem al minimizar las versiones acerca de una presunta xenofobia de sus compatriotas contra los originarios de Paraguay.

The PANGLOSS output read:

> Menem said that the argentinean did not have a fear of foreigners on the paraguayans.
>
> The argentinean president, Carlos Menem, disparaged this dominic the version on top of the existence onto its region of somewhat type of persecution against paraguayan immigrants. "About what xenophobia speak", say when the agent when he was ask for paraguayan journalists why the argentinean "irk" of the immigrant citizens of this country. Menem asserted that "does not there is no one type of difference" between argentineans and paraguayans. "Get out of argentinean and question of any of the one million paraguayan residents by my country if be or not harassed and if have somewhat type of snags in rep blica argentina", said Menem to the minimize the versions drop in a presume fear of foreigners from its countryman on the natives of Paraguay.

Frederking wrote me: "I'm not sure it's from the best version of the system, but it strikes me as the kind of quality we were getting." He concludes that "the KBMT system couldn't cover the unrestricted 'domain'" (personal communication, October 29th, 2009).

And the quality of the translation is certainly not impressive. Almost every single preposition is badly picked, from *on top of the existence* to *on the natives of Paraguay*. Some pronouns are lost rather than reconstructed (*if be or not harassed*). Bad or simply wrong translations are picked for several words (*get out of* for *vayan a*, *drop in* for *acerca de*, *agent* for *mandatario*). And some of inflections are translated incorrectly, too (*argentinean* instead of *Argentineans*, capitalised and in plural, and *did not* instead of *do not*).

To put it crudely, it would be hard to distinguish this translation from the 1966 machine translation quoted on page 32. ARPA seemed to share this dissatisfaction with the quality (White and

O'Connell, 1994), and in the course of the period 1992–94, the project was redefined to satisfy the investors:

> Originally, PANGLOSS was supposed to be a pure knowledge-based machine translation (KBMT) system implemented in a version of the interlingua architecture. The project, however, evolved toward a more eclectic approach, mostly due to the necessity to perform well during periodical external evaluations whose timing and frequency was established after the project started. . . . The project team then decided to channel some of the resources into developing an interim evaluation system which would show an immediate improvement in output quality, while in parallel continuing to develop the "mainline" KBMT system. The idea was that at a certain stage the KBMT system would supplant the interim system (Frederking et al., 1994, p. 73).

The more recent history of machine translation, however, shows that the field generally did not go in that direction. The expensive and slow construction of knowledge-heavy systems was large overtaken by the development of statistical methods that gave comparable output at a fraction of the cost. This new strategy began winning its first recognition during the period that PANGLOSS shut down (1992–96). I will return to the subject of statistical machine translation in the next section.

For now, though, I will reflect a little more on what went wrong with knowledge-based machine translation, apart from the cost. The next section thus contains some speculations about the problems with knowledge bases and inference mechanisms.

2.2.7 Why gravity drowns

All of the knowledge-based machine translation systems build in the 1980s seemed to suffer from the same flaw: They did not have knowledge enough. No matter how much money and effort was spend on building bigger databases, they never seemed to be quite big enough.

This problem had already been foreshadowed to some extent by the artificial intelligence efforts in the 1960s and 1970s. Although James Meehan had some success describing and formalising the conditions of a good narrative (cf. p. 35), he would also get bizarre output like the following:

> One day Henry Crow sat in his tree holding a piece of cheese in his mouth . . . [H]e became hungry, but he knew that he owned the cheese. He felt pretty honest with himself, so he decided not to trick himself into giving up the cheese. He wasn't trying to deceive himself, either, nor did he feel competitive with himself. But he did dominate himself, and was very familiar with himself, so he asked himself for the cheese. (Meehan, 1977, p. 92)

Meehan had done a fairly good job enumerating and describing Henry's different options and their preconditions—except for the obvious fact that none of his tricks, threats, and pleas make sense when they are directed towards himself (cf. Meehan, 1981, p. 238). A similar problem occured in another story (Meehan, 1977, p. 92):

> Henry Ant was thirsty. He walked over to the river bank where his good friend Bill Bird was sitting. Henry slipped and fell into the river. Gravity drowned.

And of course it did—having no legs or arms to swim with. Again, Meehan had formulated a rule that seemed reasonable enough in the abstract, but dubious at best in the concrete applications (cf.

Meehan, 1981, p. 218ff). The most simple, everyday facts about how the world hangs together seemed to have been left out.

All of these problems could of course be patched by adding more rules to the rules. The question is whether such a trial-and-error strategy would converge to anything, or whether it would be fundamentally flawed. If any rule might turn out to have an exception, the exceptions might have their own exceptions, and Meehan's methodology did certainly not address this issue at its root.

Although ontologies like WordNet are more systematic, they too are created from finite resources, and for every fact they contain, there is one they lack: We can learn from WordNet that `flour (noun) 1` is a substance meronym of `bread (noun) 1`, but also that neither `yeast (noun) 1`, `salt (noun) 2`, or `water (noun) 6` is. Such omissions are extremely common, both for obvious part relations like the hair on a head and for more far-fetched examples like the air in a room.

Similar deficiencies are found for the other relations, including WordNet's hypernym/hyponym tree: The `Sun (noun) 1`, for instance, is not considered a `light source (noun) 1`, and a `cow (noun) 1` is not considered a `farm animal (noun) 1`.[4] More subtle relations like ones between councilmen, violence, and demonstrators, are far beyond the scope of WordNet, and it would be useless for disambiguating sentences like (2.44 a–2.44 b) on page 33. It could seem, then, that no amount of facts would ever be facts enough.

A more recent experiment by a group at Xerox' Palo Alto Research Center—the birthplace of lexical-functional grammar—confirms this suspicion. Using the their fact extractor BRIDGE, the group attempted to complete the 2007 Real Text Entailment challenge (Dagan et al., 2006), which consisted in 410 source texts, hypotheses, and entailment values like the following:

Source text: Brigham Young University began its Jerusalem study program in 1968, first at the City Hotel in East Jerusalem.

Hypothesis: Brigham Young University was founded in 1968.

Entailment: NO

The Palo Alto group applied a cautious, high-precision methodology, guessing at a YES in only 52 out of the 410 cases. "However, we were not perfect even from this perspective," they admit and give the following example of an entailment that they got wrong (Bobrow et al., 2007, p. 21):

Source text: Girls and boys are segregated in high school during sex education class.

Hypothesis: Girls and boys are segregated in high school.

Here, there is of course no entailment—*school segregation* means constant segregation, not occasional segregation in the sense of, say, *school violence*. But since Bobrow et al. "do not yet handle generic sentences properly, our algorithm for calculating specificity produces the wrong result here" (Bobrow et al., 2007, p. 21).

Unfortunately, the distinction between the generic and the specific is both common and quite hard to grasp without a keen sense of plausibility and normality. The article concludes:

---

[4]The facts actually recorded are problematic, too. We can, for instance, learn from WordNet that a `penguin (noun) 1` is a `bird (noun) 1` and hence has feathers, `feather (noun) 1`. Similarly, the database implies that a `tent (noun) 1` has a `foundation (noun) 3`, and that a `secondhand_car (noun) 1` as well as a `Model_T (noun) 1` has an `air_bag (noun) 1`.

These problems stem directly from the model of cognition and categorisation that WordNet is build on (cf. for instance Margolis and Laurence, 1999), and they will look different in view of the theory in chapter 3. I will return to the issue in particularly sections 4.1 and 4.1.

> For example, the most likely interpretation of *Counselors are available* is episodic: SOME counselors are available. But *Experts are highly paid* is weighted towards a generic reading: MOST IF NOT ALL experts get a good salary.
>
> These examples are indicative of the subtlety of analysis necessary for high precision textual inference. (Bobrow et al., 2007, p. 21)

They indeed are, and they are also indicative of how far knowledge-based language processing is from doing inference anything like what a human being does, even after 40 years of research. Of course, we might still hope that, within another 40 years, a knowledge-based system would be able to understand sentences like *Counselors are available* and *Experts are highly paid*. But there are reasons deeper than impatience for considering this a vain hope. These reasons will receive considerable attention in chapter 3, but for now, I will turn to the topic of statistical machine translation.

2.3 Statistical methods

2.3.1 Codes, noise, and backwards inference

In his work on cryptography, Claude E. Shannon had noted that the statistical properties of a language like German could be used for code-breaking (Shannon, 1949).[5] German text has a characteristic profile of letter frequencies. If a German message is encrypted by substituting code words for the letters, the letter frequencies will reappear as code word frequencies. For instance, if 5.7% of a German text consists of the letter *r*, then some code word will occur in the cryptogram with frequency 5.7%.

The profile of an encrypted message thus hints at probable decryption keys. Such keys can be evaluated by doing a backwards substitution—putting *r*'s back in place of code words—to see if that actually yields readable, German prose (Shannon, 1949, §11).

In other words, when a code-breaker encounters a cryptogram E, he or she should consider different encryption keys that could have masked a message M as the cryptogram E. For any such key, the cryptogram E will be attributed to a different original message M. If the code-breaker decodes E according to some reasonable encryption scheme and the decoding yields prose rather than jabberwocky, then the reconstructed text probably is the original source message.

Shannon thought of this decoding process in terms of Bayes' rule (Shannon, 1949, §10):

$$P(M|E) = \frac{P(M)P(E|M)}{P(E)}.$$

Interpreted as a cryptographical equation, Bayes' rule elaborates $P(M|E)$, the probability that a text M is the source message behind a cryptogram E. The right-hand side of the equation expresses this conditional probability in terms of the "Germanness" of the message, $P(M)$, the *a priori* cryptogram probability, $P(E)$, and the probability that some encryption scheme would encrypt M as E, $P(E|M)$.

The "Germanness" or "Englishness" of a string of words or characters was a topic that Shannon would spend more energy on in his writings on information theory (Shannon, 1948; Weaver and

[5]The contents of his paper "Communication theory of secrecy systems" was written during the war, but was classified until 1949 when it appeared in the *Bell System Technical Journal*.

Shannon, 1949). English text, he noted, is not a constant flow of characters with equal status, but partly predictable. Borrowing a term from thermodynamics, Shannon called the room for actual information the "entropy" of a code or language. For English, he estimated that around 50% of any given text was determined by structure rather than content (Shannon, 1948, §7).

Knowing little about grammar and nothing about structuralist linguistics, Shannon mainly thought of this predictability in terms of n-gram frequencies. To illustrate this predictability, he produced a series of random texts with (a) letter frequencies like English, (b) bigram frequencies like English, and (c) trigram frequencies like English (Shannon, 1948, §3):

(2.57) (a) ocro hli rgwr nmielwis eu ll nbnesebya th eei alhenhttpa oobttva nah

(b) on ie antsoutinys are t inctore st be s deamy achin d ilonasive

(c) pondenome of demonstures of the reptagin is regoactiona of cre

He commented:

> The resemblance to ordinary English text increases quite noticeably at each of the above steps. ... It appears then that a sufficiently complex stochastic process will give a satisfactory representation of a discrete source (Shannon, 1948, §3).

While this claim was largely discredited by the later development of linguistics, Shannon set the stage for a new way of looking at language, namely as an information channel characterised by its frequency profile. The idea of "decoding" a message obtained from such a channel consequently seemed quite natural when Shannon's colleague Warren Weaver began toying with the idea in 1949 (cf. p. 7).

Over the course of the last two decades, this idea has been expanded into an actual machine translation programme. In the following subsections, I introduce these recent inventions.

2.3.2 In French code

In the latter part of the 1980s, neural networks and machine learning had been hot topics in artificial intelligence and computer science. For tasks like pattern recognition, "trained" programmes had begun to outperform highly controlled, rule-based programmes.[6] By then, the Canadian government had also published their entire record parliamentary proceedings machine-readable form—thousands of pages of hand-translated, parallel text in French and English (cf. table 2.5, p. 46).

With such methods and resources at hand, it did not seem completely out of the question that a computer programme might be able to learn about translation on its own, rather than be instructed in translation. In order for these resources to be turned into a translation algorithm, however, a model of the goal and the training situation had to be created.

In 1990 and 1993, a research group at IBM published two papers on what they called "statistical machine translation" (Brown et al., 1990, 1993). The papers outlined a method for creating a set of useful statistics and described how these statistics could be turned into a translation algorithm.

The IBM group took Bayes' rule as their theoretical starting point, but unlike Claude Shannon, they explicitly interpreted it in terms of translation and grammar:

$$P(e|f) = \frac{P(f|e)P(e)}{P(f)}.$$

[6]The first couple of issues of the *Neural Networks* (Grossberg, 1988) contain many examples and show the sense of expectation that was in the air at the time.

1. `We acknowledge the government's desire to create links, forums for interaction, for exchange, for focusing researchers' efforts.`

   `Nous sommes reconnaissants au gouvernement de vouloir créer des lieux de maillage, des lieux d'interaction, des lieux d'échanges, des lieux de densité entre les chercheurs.`

2. `We agree that this is the path modern research needs to take, but we believe this bill has a certain number of flaws, and have sought to improve it.`

   `Nous acceptons que la façon moderne de faire de la recherche passe par cette voie. Mais nous croyons que le projet de loi est caractérisé par un certain nombre de faiblesses que nous avons cherché à bonifier.`

3. `We introduced about thirty of them, each one more relevant than the last, and these were amendments which witnesses had called for.`

   `Nous en avons déposé une trentaine, tous plus pertinents les uns que les autres, des amendements qui étaient demandés par les témoins.`

4. `Unfortunately, the government turned a deaf ear to them.`

   `Malheureusement, le gouvernement a fait la sourde oreille.`

Table 2.5: A small excerpt from the proceedings of the French Parliament from March 28th, 2000.

The conditional probabilities $P(e|f)$ and $P(f|e)$ they thought of as the probabilities that a French string f would translate as the English string e and *vice versa*. $P(e)$ and $P(f)$ represented the grammaticalness of the two sentences.

Thus, when a decoder encounters a French string f, the model prescribes that it should search for the English string e that maximises $P(e|f)$. According to Bayes' rule, this is theoretically speaking the same as searching for an e that maximises $\frac{P(f|e)P(e)}{P(f)}$. Since the French sentence f would be held constant during such a search, the e that maximises this expression should in theory be the same as the e that maximises

$$P(f|e)P(e).$$

The problem of "decoding" a French string f was thus restated from a task of maximising $P(e|f)$ to maximising $P(f|e)P(e)$. Instead of searching for a likely translation of f, the decoder should search for a string e fulfilling two conditions: e should translate into something like f in a English-to-French translation, and it should look like a proper English sentence.

In principle, the IBM group had won nothing by reformulating the problem like that: They just had to estimate the quality of English-to-French translation instead of French-to-English. Things had even been complicated by the introduction of $P(e)$, the estimate of the grammaticalness of e. In a world of perfect probability estimates, they would just have found a more complicated way of computing $P(e|f)$.

In a world of less than perfect estimates, however, there was a good reason for preferring two distinct factors over one: If they constructed a translation algorithm based on a flawed estimate of the translation probability $P(e|f)$, it would inevitably produce flawed translations. But by splitting the process up into two distinct estimates, the merits of $P(f|e)$ and $P(e)$ might, with a little luck, complement each other and smooth out the flaws each estimate would have in isolation. This is an important point, so I want to elaborate it a little.

Imagine a Martian who speaks neither French nor English, but owns an English–French dictionary (cf. also Knight, 1997a, pp. 83–88). Such a Martian might try to guess whether f is a translation of e by the following method: First, he looks up all the dictionary senses of every word in e. Then, he counts the words in f that match one of these dictionary senses. If this number seems high, he guesses that e actually translates as f.

That is of course a problematic method. The English word *sheet* might have been used in the sense of *feuille* in a sentence, but incidentally match the French word *plaque* and thus give a false impression of relatedness. Or f might contain an important word like *ne* or *jamais* that nothing in e accounts for—and since the Martian knows no French, this will seem as significant or insignificant as any random word.

If the Martian was asked whether *une carte de la région* was a translation of *a map of the region*, *a card of the region*, or *a ticket of the region*, he would not have a qualified answer. If he were to translate the phrase himself, his best bet would be to flip through the pages of his dictionary, looking for words that translated into *une*, *carte*, *de*, etc. This might result in any of the three English constructions.

But the Martian might have a friend who understands a little English. If these two aliens should decide to collaboratively translate a French sentence, they could use the following recipe: First, the Martian with the dictionary begins to search for English words that translate into some word in the French sentence and writes down what he finds:

	une	carte	de	la	région
	a	card	of	...	
(2.58)	an	map			
	one	ticket			

When he is some way into this process, his friend begins to rule out word senses that fit badly into the whole:

	une	carte	de	la	région
	a	card ·/·	of	the	region
(2.59)	an ·/·	map	from		area
	one ·/·	ticket ·/·			section ·/·

When all the French words are accounted for by some English word that fits reasonably with its context, there are hopefully just a small number of ways to combine them into a plausible sentence:

	une	carte	de	la	région
	a	card ·/·	of	the	region
(2.60)	an ·/·	map	from		area
	one ·/·	ticket ·/·			section ·/·

This combination of English words is then accepted as the translation of the French sentence.

This little recipe is not completely different from the translation algorithm proposed by the IBM group. It does, however, differ on some accounts: First, it does not deal with change of word order or sentence length at all. Second, in the IBM methodology, the Martian's dictionary is not a dictionary in the traditional sense, but a battery of probability tables constructed automatically from a bilingual corpus. And third, the friend is actually the statistical profile of the English language—that is, a table of n-gram probabilities.

More formally, the statistical machine translation method proposed by the IBM group assumes three different modules:

1. A translation model, used as an estimate of $P(f|e)$
2. A language model, used as an estimate of $P(e)$
3. A search heuristic or "decoder," used to limit and guide the search through the vast set of English words sequences

In the previous analogy, these correspond to (1) the Martian with the dictionary, (2) his anglophone friend, and (3) the recipe I described. In the following subsections, I will explain the contents of these modules in a little more detail.

2.3.3 The dictionary that writes itself

In parallel texts like the example in table 2.5, words that we intuitively think of as translations of each other have a tendency to cooccur. For instance, if an English sentence contains the word *bill*, there is a good chance that the corresponding French sentence contains the word *projet*. Of course, the French sentence may contain other words like *la*, *de*, *la*, *que*, etc., but when we look at more sentences, *bill* usually turns out to cooccur with *projet* more persistently than other French words.

Because of this persistency, it is possible to construct a table of word translations automatically. This is the idea behind the translation model that Brown et al. referred to as "Model 2" (Brown et al., 1993, §4.2).[7] Since Model 2 is a translation model, it is supposed to give some kind of model of how English sentences account for French sentences. It does that in terms of a translation table t and an alignment table a.

The translation contains the probabilities of word translations like

(2.61) $t(le\,|\,the)$ = 61.0%, $t(la\,|\,the)$ = 17.8%, $t(l'\,|\,the)$ = 8.3%, etc.

The alignment table contains probabilities of connections—for instance, the probability that the word on place 1 in a French sentence is accounted for by the word on place 3 in an English sentence:

(2.62)
I am sleeping
↗ 33.3%
Je dor

[7] There is, of course, also a Model 1. It consists of a translation table constructed by counting how often pairs of words cooccur, without introducing alignments as explained later in this subsection. For some purposes, Model 1 is a convenient starting point, but it is not strictly necessary, and I will consequently not discuss it further.

The probabilities of these connections are conditioned on the sentence lengths. They consequently have the form $P(i \mid j, |e|, |f|)$, where j is a position in the French sentence, i is a position in the English, and $|e|$ and $|f|$ are the lengths. Every French word is attributed to exactly one English word.[8]

Such connection probabilities can also be used to compute probabilities of entire sentence alignments such as

(2.63)

I am sleeping	I am sleeping	I am sleeping	
Je dor ,	Je dor ,	Je dor ,	etc.

Assuming that the individual connections are independent, we can estimate the probability of such a sentence alignment by the product of the connection probabilities. In the case of the leftmost alignment above, for instance, this joint probability would be $a(1 \mid 1, 3, 2) \times a(3 \mid 2, 3, 2)$.

Brown et al. created these tables by an iterative process in which the alignment table was used to improve the estimate of the translation table and *vice versa*. More specifically, they began by setting all values of a and t to the same number. Then they used a to estimate t, then t to estimate a, then a to estimate t, then t to estimate a, etc. This method is a special case of the so-called "expectation–maxization algorithm", which is used in applied statistics when parameter estimation is needed from an incomplete data set (Baum, 1972; Dempster et al., 1977). As the number of reestimations by this method increases, both the parameter set and the augmented data converge to an equilibrium.

Let me show this in a little more detail. Consider the following Dutch–English "corpus":

(2.64) (a) Chapter one | Eerste hoofdstuk

(b) Chapter two | Tweede hoofdstuk

To train a Model 2 from this corpus, we run through every Dutch target word in the corpus: *Eerste*, *hoofdstuk*, *Tweede*, *hoodfstuk*. For all these words, we consider every English source word that might be aligned to the Dutch word—for instance *Chapter* and *one*. in the case of *Eerste*.

For all such pairs, we note how probable the connection seems in light of the translation probability, and how probable the translation seems in light of the connection probability. This weighted count results in two new probability tables. We then start over and do a new count on the basis of these updated tables (cf. table 2.6 on page 50).

In this specific microcorpus, *hoofdstuk* cooccurs with *Chapter*. The training therefore gradually assigns a higher value to $t(\textit{hoofdstuk}|\textit{Chapter})$. This slowly causes word order reversion to be the preferred alignment. That, in its turn, assigns a higher probability to the slightly idiosyncratic translations *Eerste*|*one* and *Tweede*|*two*, as well as *hoofdstuk*|*Chapter*. As the algorithm is repeated, these effects are reinforced.

In a slightly more realistic setting, results like the ones in table 2.7 on page 53 begin to appear: The German word *Beifall* is often attributed to *Applause* and *applause*, but also to *Loud* and *sustained*, which often occur in the same contexts; the slash (/) is attributed to both *HIV* and *AIDS* because of the frequent construction *HIV/AIDS*; the first person inflection *bin* ("am") is attributed to seemingly

[8] In order to account for French words with no obvious source, Brown et al. inserted an invisible 0th word in every English sentence in their corpus. French words otherwise unaccounted for could then be attributed to this "null word" (Brown et al., 1990, p. 81). This has some consequences for the number of words that a decoded sentence will be equipped with, but it would require to much space to go into this issue.

Trans.	*Eerste*	*hoofdstuk*	*Tweede*
Chapter	33.3%	33.3%	33.3%
one	33.3%	33.3%	33.3%
two	33.3%	33.3%	33.3%

Align.	Prob.
↑↖	25.0%
↑↑	25.0%
⤧	25.0%
↗↑	25.0%

Trans.	*Eerste*	*hoofdstuk*	*Tweede*
Chapter	50.0%	33.3%	50.0%
one	50.0%	25.0%	0.0%
two	0.0%	25.0%	50.0%

Align.	Prob.
↑↖	25.0%
↑↑	25.0%
⤧	25.0%
↗↑	25.0%

Trans.	*Eerste*	*hoofdstuk*	*Tweede*
Chapter	42.9%	60.0%	42.9%
one	57.1%	20.0%	0.0%
two	0.0%	20.0%	57.1%

Align.	Prob.
↑↖	33.3%
↑↑	16.7%
⤧	33.3%
↗↑	16.7%

Trans.	*Eerste*	*hoofdstuk*	*Tweede*
Chapter	36.0%	69.2%	36.0%
one	64.0%	15.4%	0.0%
two	0.0%	15.4%	64.0%

Align.	Prob.
↑↖	30.9%
↑↑	10.8%
⤧	43.2%
↗↑	15.1%

Trans.	*Eerste*	*hoofdstuk*	*Tweede*
Chapter	29.8%	77.2%	29.8%
one	70.2%	11.4%	0.0%
two	0.0%	11.4%	70.2%

Align.	Prob.
↑↖	27.6%
↑↑	6.6%
⤧	53.0%
↗↑	12.7%

Trans.	*Eerste*	*hoofdstuk*	*Tweede*
Chapter	24.4%	83.8%	24.4%
one	75.6%	8.1%	0.0%
two	0.0%	8.1%	75.6%

Align.	Prob.
↑↖	24.0%
↑↑	3.9%
⤧	62.1%
↗↑	10.1%

Table 2.6: A number of iterations in a Model 2 training. The parameters gradually converge to an equilibrium in which *one*, for instance, is considered a likely translation of *Eerste* ("First"). The alignment ⤧, reversing the word order, also gradually obtains a higher score.

unrelated words like *happy*, *Great*, and *dismayed*, because *bin* mostly occurs when a speaker in the parliament expresses a personal opinion (Brown et al., 1990, p. 83).

2.3.4 Guessing and reconstructing

The alignment and translation tables which emerge from Model 2 training define a translation model: Two sentences f and e are semantically close if you can draw a set of arrows from the words in f to the words in e so that (1) every word in f connects to some word in e, (2) the arrows are probable alignments, and (3) the connected words are probable translations of each other.

A simple language model can be build from a corpus by counting every n-gram in the corpus for, say, $n = 3$ or $n = 2$. The "Englishness" of a sentence can then be estimated as the product of the frequencies of the n-grams it contains. For instance, if $n = 2$, the probability $P($*it is raining*$)$ would be estimated by multiplying the frequencies of `("it", "is")` and `("is", "raining")`, as they were computed from the corpus.

The contours of a simple decoding algorithm thus begins to emerge. Given a French sentence f, a decoder should apply the following recipe (cf. Brown et al., 1993, pp. 275 and 304):

1. Pick a handful of sentence lengths $|e|$ that seem probable in light of the length $|f|$. In Brown et al.'s algorithm, all lengths are assigned equal probability (Brown et al., 1993, p. 268), but this is not terribly important.

2. For every length hypothesis, pick a handful of slots in e that some slot in f is likely to be attributed to. If you had 10 lengths hypotheses and picked 10 alignments hypotheses for each of those, you should now have 100 length/alignment hypotheses.

3. Drop the length/alignment hypotheses that are most unlikely. You can determine the probability of a length/alignment hypothesis by multiplying the length probability and the alignment probability.

4. For every slot in e which is connected to a French word f_j by some length/alignment hypothesis, pick a handful of probable translations of f_j. This will give you a number of length/alignment/translation hypotheses. In all of these, exactly one slot in e is filled by an actual word.

5. Drop the length/alignment/translation hypotheses that are most unlikely.

6. For every length/alignment/translation hypothesis, pick a new slot in f and align it to a handful of slots in e. This gives you a new set of alignment hypotheses.

7. Drop the least probable alignment hypotheses.

8. For every slot connected to a word f_j by an alignment hypothesis, pick a handful of translations of f_j.

9. Drop the least probable alignment/translation hypotheses.

10. Repeat steps 6–9, gradually elaborating the hypotheses. As soon as the hypothesised e's contains adjacent words, take the probability of the n-gram into account when evaluating the hypotheses.

11. When you have repeated steps 6–9 until all words in f have been translated ($|f| - 1$ times), pick the most probable remaining hypothesis ones and return that as the translation.

Through this constant oscillation back and forth between elaboration and pruning of the search tree, the decoder is able to look at promising parts of the search space without facing an unmanegable number of hypotheses.

2.3.5 Beyond words

There is no single, obvious way to use the alignment and translation tables that come out of Model 2 training. The translation algorithm described in section 2.3.4 will produce output, but has drawbacks too severe to be taken seriously—for instance, it always decreases sentence length during translation.

Brown et al. suggested several ways to use the tables as input for further training and more advanced decoding algorithms. Their "Model 3" was build around the concept of "fertility" rather than alignment (Brown et al., 1993, p. 275). Fertile words are English words like *not* that tend to account for more than one French word. The probability of a particular set of connections between f and e would thus be different depending on the actual words in e. The connection probabilities would further be modified by the "distortion" they postulated, that is, the distance they claimed that an English word had to be moved when translated.

Such a model cannot be trained efficiently by a simple count the same way that Brown et al.'s Model 2 could. They consequently had to use the parameter estimates from a Model 2 training as a starting point and then gradually adjust the new parameters until they seemed to reach an equilibrium. These new parameters could then be used for a search heuristic with a refined translation probability estimate (Brown et al., 1993, §4.4).

Brown et al. further refined their translation models, resulting in a Model 4 and 5. However, these models later turned out to give quite little improvement in translation quality in return for their large computational cost (Koehn et al., 2003). They have consequently been applied much less than the other models.

All of the IBM models had the feature in common that they could only attribute a French word to a single English word. They did, in that sense, give rise to word-for-word translation systems and relied on the language models to clean up the problems that such a window translation approach created. In a 1999 paper, Franz Josef Och, Christoph Tillmann, and Hermann Ney suggested that this simplification could be overcome by using two dictionaries instead of one (Och et al., 1999).

That is, instead of just aligning the French words in the parallel corpus to the English words, they would turn the corpus upside down and do the entire training once again to obtain yet another dictionary. This would allow them to connect the words in a sentence pair in a many-to-many fashion instead of only a many-to-one (cf. figures 2.6 and 2.7, p. 53).

Consequently, they could also begin to actually look at standard phrases as more than just a biproduct of high translation probabilities: Since they had an estimate of the probability of any alignment between two sentences, they could also estimate the probability of seeing certain clusters of words paired with other clusters of words. The candidates for such phrases were found by scanning symmetric alignment matrices like the ones in figures 2.6 and 2.7. I will, however, not go further into the details of the techniques behind such phrase-based machine translation.

Paradoxically, around the same time that alternatives to the IBM models 4 and 5 began to appear, those models started to the attention of a wider audience. This might have been due to several factors.

/		*Beifall*		*bin*		*Krieg*	
/	76.7%	*Applause*	33.3%	*happy*	50.0%	*War*	33.3%
96	25.0%	*applause*	25.0%	*Great*	10.0%	*unrest*	7.7%
HIV	16.7%	*Loud*	25.0%	*dismayed*	9.1%	*civil*	7.7%
AIDS	16.7%	*sustained*	16.7%	*gates*	8.3%	*1986*	6.7%

Table 2.7: Selected translation probabilities after 25 iterations of Model 2 training. The model was trained on approximately 3,000 short sentences from English and German translations of European Parliament proceedings.

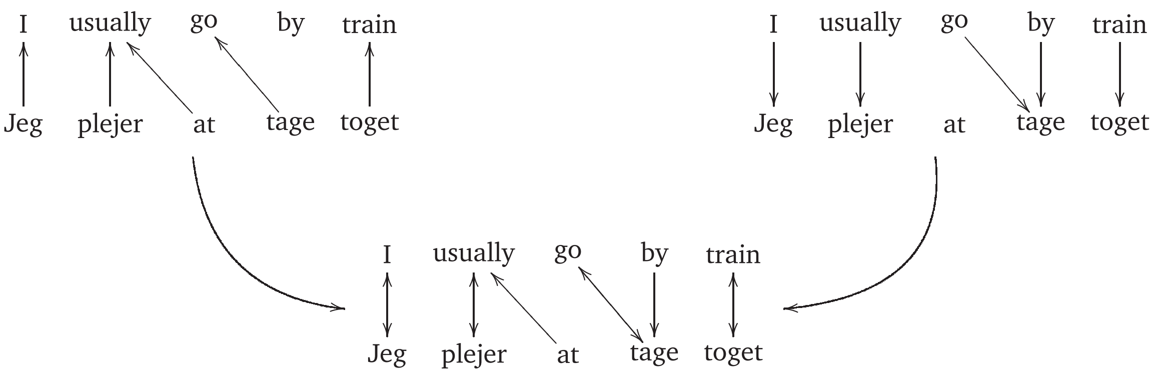


Figure 2.6: The union of a Danish-to-English and an English-to-Danish alignment.

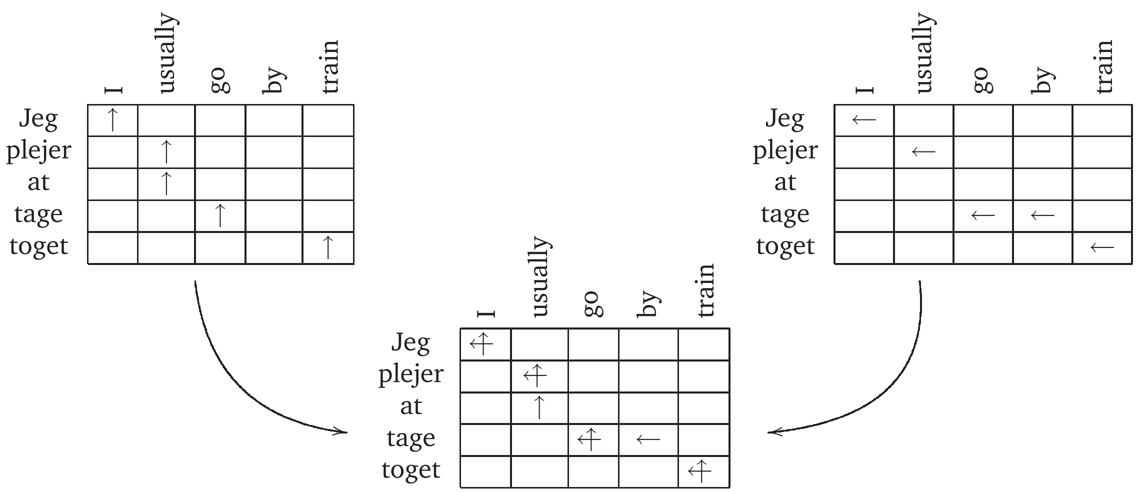


Figure 2.7: The union of the same alignments as in figure 2.6, represented as matrices.

For instance, during a workshop at John Hopkins University in 1999, a group consisting of Yaser Al-Onaizan, Kevin Knight, Franz Josef Och, and others constructed a statistical machine translation system. As a part of the process, they also created a software package for training translation models, called GIZA (Al-Onaizan et al., 1999). To demonstrate the efficiency of the method, they ended the workshop by constructing yet another statistical machine translation system in a single day (Al-Onaizan et al., 1999, p. 2).

For that same workshop, Kevin Knight also wrote a informal introduction to statistical machine translation, *A Statistical MT Tutorial Workbook* (Knight, 1999). The booklet unpacked the content of Brown et al.'s somewhat convoluted presentation and was circulated at least enough become the basis of some of Daniel Jurafsky and James Martin's textbook introduction to the subject (Jurafsky and Martin, 2008, pp. 935–39).

Several other tools were produced in the following years by different machine translation researchers, including Philipp Koehn's decoder Pharao (Koehn, 2004a,b) and the machine translation system MOSES (Koehn et al., 2007). Google also began to show interest in the topic, and in 2001, they introduced their first statistical machine translation system. In 2004, Franz Josef Och joined the company, and he is currently head of their machine translation department.

2.3.6 "They are not equal in panick"

Bad output from statistical machine translation systems is often defended with the claim that the systems would do better for texts more like the training corpus, just like knowledge-based machine translation used to be defended by the claim that it would do better when it had lexical coverage.

There is some truth to this claim. Statistical machine translations do indeed vary quite a lot in quality, and the quality is often dependent on both the language pair and the topic of the text. Texts about politics generally do better in most systems, and French-to-English translations typically yield better results than, for instance, Japanese-to-English.

The best statistical machine translation has to offer is thus translations of texts like the following:

> Janvier 2009, lendemain de la Saint-Sylvestre. Une vague de froid provenant de l'Arctique s'abat sur l'Europe. A 7 heures du matin, tandis que les Européens sont encore mal remis des libations de la veille, Gazprom annonce que la livraison de gaz a été interrompue. Les tensions entre Moscou et Kiev sur le prix du gaz naturel ont amené la Russie à fermer le gazoduc qui traverse l'Ukraine et qui dessert aussi l'Europe centrale et orientale. La coupure durera 20 jours, empêchant des pays comme la Slovaquie, la Bulgarie ou la Moldavie d'alimenter en gaz une partie de leur population, alors même que l'hiver est particulièrement rude.

Philip Koehn has made a version of the MOSES system, trained on the European Parliament proceedings, publicly available on the website `http://demo.statmt.org/`. On that system, the translation of the text above reads (produced December 10th, 2009):

> January 2009, after the Eve. A wave of cold from the Arctic beats down on Europe. Has 7 o'clock in the morning, while Europeans are still badly placed on the eve of libations Gazprom announcement that the supply of gas has been suspended. The tension between Moscow and Kiev on the price of gas has led Russia to close the gas pipeline across Ukraine and serves as central and eastern Europe. The last cut 20 days, preventing countries such as Slovakia, Bulgaria or Moldova to nurture gas part of their population, while the winter is particularly harsh.

In spite of the fact that the text contains fairly long, reasonable strings, the quality of this translation is poor by any human standard. None of the sentences are proper English, and several quite severe errors have crept in: The text seems to suggest that Russia *serves as central and Eastern Europe*; the embargo that lasted 20 days has become *the last cut*; and the populations of Slovakia, Bulgaria, and Moldova have apparently grown a *gas part*. Several words from the French text have also mysteriously disappeared, and obviously wrong lexical choices like *Has 7 o'clock* have been made.

In other words, the performance of the system is not impressive when we look at its actual output. The situation gets worse if we change the subject and the language pair slightly. Consider for instance the following German text:

> Wenn auf dem Motherboard Montagelöcher vorhanden sind, diese aber nicht in der gleichen Richtung verlaufen, wie die auf dem Boden und keine Slots zur Anbringung von vorhandenen Abstandhaltern vorhanden sind, geraten Sie nicht gleich in Panick; Sie können immer noch die Abstandhalter mit den Montagelöchern verbinden. Entfernen Sie einfach die untere Seite der Abstandhalter (es kann etwas schwierig sein, den zu entfernen; seien Sie daher vorsichtig). So können Sie das Motherboard immer noch auf dem Boden anbringen und brauchen sich keine Gedanken über einen Kurzschluss zu machen.

The MOSES output, also based on the European Parliament proceedings, reads:

> If, on the motherboard montagelöcher are there, but not in the same direction, such as those on the ground and not to use of existing abstandhaltern slots available that they are not equal in panick; you can still with the montagelöchern combine the abstandhalter. They simply remove the lower side of the abstandhalter (there can be a little difficult to remove; they are therefore cautiously). So you can motherboard still on the ground and need not worry about a to dispel.

I take it that no one would consider this a good translation, regardless of its BLEU score.

Neither these examples nor the Google Translate examples in the introduction give us much reason to expect that statistical machine translation is about to reach the quality of ordinary translation anytime soon. In the next subsection, I will give a tentative explanation for the errors we see in examples like these.

2.3.7 Where garbage goes

The machine translation systems of the late 1980s needed enormous sets of rules in order to function. Constructions like *live out*, *live for*, and *live by* had to be distinguished from *go out*, *go for*, and *go by*, and every one of them had to be individually described and translated. Statistical machine translation reduced the length of process from years to hours.

But it is not completely unproblematic to throw out everything once knew about language out the window. Like the systems of the early 1950s, statistical machine translation relies heavily on local contexts and is highly unreliable when decisions require information from more than just a few neighbouring words. This will necessarily cause errors.

For instance, the Danish adjective *let* may mean either "easy" (as opposed to "hard") or "light" (as opposed to "heavy" or "substantial"). A translation model would assign each of the translations *let|easy* and *let|light* a probability. Since *let* is used slightly more often in the sense of "easy," we would probably have $t(let|easy) > t(let|light)$.

The translation probabilities would therefore dictate that *an easy closet* be considered before *a light closet* as a translation of *et let skab*. The translation model would have to rely on the language model to discard the incorrect phrase *an easy closet* by penalising the bigram `("easy", "closet")`.

The problem, however, is that simplistic language models do not always put two and two together the way we hope they do. If nothing in the immediate context disagrees with the choice of *easy*, it is accepted. This might happen in sentences like

(2.65) Den nye seng er ikke så let som den ser ud.

"The new bed is not as light/easy as I it looks."

Here, the local context might even support the wrong choice: The phrase *not as easy as it looks* is about 17.5 times more common than the phrase *not as light as it looks* (according to a web search).

In other cases, the information relevant to the decision might lie several sentences back:

(2.66) Vi lagde flere sten i urtepotten. Det så ikke ud til at hjælpe. Den var stadig for let.

"We added more pebbles to the flowerpot. It didn't seem to help. It was still to light/easy."

Similarly, the English construction *carry* usually translates as *dragen* ("move while supporting") in Dutch. However, the most probable translation of *carry out* would be *uitvoeren* ("execute, perform"). But both decisions can be reversed indefinitely by the addition of more context:

(2.67) (a) I carried... (*dragen*)

(b) I carried out... (*uitvoeren*)

(c) I carried out the garbage... (*dragen*)

(d) I carried out the garbage collection... (*uitvoeren*)

(e) I carried out the garbage collection. She handled the record collection... (*dragen*)

Similar examples could be given for other classical problems like attachment problems, pronoun resolution, and the like. The differences in meaning may be as big as between terminating a presidency and terminating a president.

As I noted earlier, there is no natural boundary on how much context it takes to make such decisions. In one classical experiment from cognitive psychology (Bransford and Johnson, 1972, p. 722) this was demonstrated by the following text:

> The procedure is actually quite simple. First you arrange things into different groups depending on their makeup. Of course, one pile may be sufficient depending on how much there is to do. If you have to go somewhere else due to lack of facilities that is the next step, otherwise you are pretty well set. It is important not to overdo any particular endeavor. That is, it is better to do too few things at once than too many. In the short run this may not seem important, but complications from doing too many can easily arise.

In the last sentence, *do* is used in a vague, generic sense. If we were to translate it into a language where such vague verbs were not available (for instance, Danish or German), we would have to make a decision regarding the meaning of the text. It would take little more than a single occurrence of the word *laundry* before the first sentence to trigger a series of decisions about the meaning of the word *do* as well as the ambiguous words, *procedure*, *arrange*, *groups*, *makeup*, *pile*, *facilities*, etc.

The standard excuse that more and better training data is called for does not help here. The decoding modules in standard statistical machine translation systems typically translate on a sentence-for-sentence basis, so anything that takes place across a period might as well take place on another planet. The size of this window can be expanded to two or possibly three sentences on a fast computer, but the search space explodes if much more is taken into account.

Brown et al. gloss over this problem with the excuse that translation cannot be expected to be absolutely perfect. In the 1990 paper, they refer to Marcel Proust's *A la Recherche du Temps Perdu*, which begins and ends with the same French word, separated by some thousand pages of text. A human translator might notice this deliberate circularity and preserve it in the English version, but with respect to machine translation, they add: "We, of course, do not hope to reach these pinnacles of the translator's art" (Brown et al., 1990, p. 79).

If the only problems statistical machine translation had were of such extreme subtlety, it would indeed have been an unqualified success. But as we have seen, it does not even see the difference between *it can be a little difficult* and *there can be little difficulty*. To pretend that such mistakes would only interest a few eccentric Proust scholars is, at best, a case of denial.

This situation is somewhat like that machine translation faced around 1950. In his Rockefeller Foundation memo, Warren Weaver argued that anything could be translated correctly using an n-gram model if n was large enough. "The formal truth of this statement becomes clear when one mentions that the middle word of a whole article or a whole book is unambiguous if one has read the whole article or book," he stated (Weaver, 1955, p. 21). Characteristically, he forgot to mention that for such a large n, the n-gram table would contain more phrases than there are particles in the visible universe. The possibility that translation might not be a matter of phrase tables did not enter the discussion.

Statistical machine translation is currently running up against a similar wall, blaming the problems on the data rather than the method. After all, all the information our systems need must be down there, in the text; the formal truth of this statement is shown by the fact that human beings can and do translate. And what else would human beings base their decisions on, if not the text?

Chapter 3

Language and philosophy

In the previous chapter, I looked at how machine translation is done and what kind of errors it produces. This gave some reason for pessimism with respect to the possibility of doing translation automatically. In this chapter, I justify that tentative conclusion by some more general and independently motivated arguments. These arguments will reveal that the problems of machine translation are not specialist issues having to do with technicalities, but are caused by deeper reasons having to do with the human condition in a broader sense.

I order to tell that alternative story, I need to refer to two schools of thought from 20th century philosophy, pragmatism and phenomenology. In this text, I use Ludwig Wittgenstein and Martin Heidegger as representatives of these two schools.

This is not because they are the first, the clearest, or the only philosophers I could have used, but they are the standard references and the two largest figures in 20th century philosophy. The argument I here propose is thus not an esoteric secret I have dug up from the archives. Quite on the contrary, it has been well-known and taught at universities worldwide throughout the history of machine translation.

Said very abstractly, the conclusion I draw from these theories is that linguistic meaning is derived from the the role things and people play in everyday life, but that these roles are neither based on facts nor rational decision. Since formalisation is a kind of rationalisation, it necessarily gets the nature of language wrong.

3.1 Objectivist semantics

Before I go into the substance of Wittgenstein and Heidegger's philosophies, I want to give a brief sketch of the traditional philosophical picture that they were arguing against. Besides clarifying their positions, it will also bring out the fact that machine translation was not born in a theoretical vacuum—rather, it drew heavily on some preconceptions about language which had already been in circulation long before the 1940s. Following Lakoff and Langacker, among others, I will refer to this set of ideas as an "objectivist" philosophy (Lakoff, 1990, ch. 11 and Langacker, 2002, ch. 11).

Some of the assumptions of objectivist philosophy were shared among professional philosophers and linguists alike in the first half of the 20th century. These can be summarised in the following points:

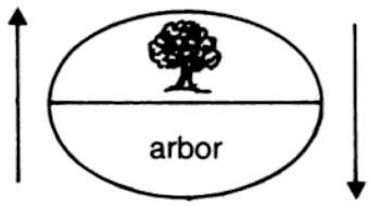


Figure 3.1: Saussure' picture of a sign (taken from de Saussure, 1986, p. 67). In this case, the sign connects the French word *arbor* ("tree," the signifier) and the general concept or thought of a tree (the signified). Such couplings are assumed to be stored in our heads: "Language exists ... in the brain of each member of a community, almost like a dictionary of which identical copies have been distributed to each individual." (de Saussure, 1966, p. 19)

- Signs are, in Ferdinand de Saussure's words, bonds between "ideas and sounds" ((de Saussure, 1966, p. 111); cf. figure 3.1).
- Sentences are thus the tangible form of abstract judgements. As Gottlob Frege says: "The thought ... clothes itself in the material garment of a sentence and thereby becomes comprehensible to us" (Frege, 1980, p. 292).
- If a word is used consistently at all, it has a definable meaning. Leonard Bloomfield explains: "We can define the names of minerals, for example, in terms of chemistry and mineralogy, ... but we have no precise way of defining words like *love* or *hate*, which concern situations that have not been accurately classified" (Bloomfield, 1935, p. 139, §9.1).
- Hence, the meaning of a sign can be split into objective content and subjective noise (cf. for instance Russell, 1923). Frege comments: "Subjective ideas are often demonstrably different in different men, objective ideas are the same for all. ... The distinction here drawn stands or falls with that between psychology and logic. If only these themselves were to be kept always rigidly distinct!" (Frege, 1980, p. 37, footnote).
- Signs are unmotivated: "The bond between the signifier and the signified is arbitrary" (de Saussure, 1966, p. 67). Even when there is a motivation (Saussure's example is a loanword), the motivation does not belong to the sign as such, and "the more rigidly they are kept apart, the better it will be" (de Saussure, 1966, p. 22).
- Considered as a whole, language is "a system of signs that express ideas" (de Saussure, 1966, p. 16). This systematicity implies that it can be thought of in terms of rules. Rudolph Carnap spells that out in a 1931 article: "A language consists of a vocabulary and a syntax, i.e. a set of words which have meanings and rules of sentence formation" (Carnap, 1996, p. 11).
- Observable speech is an application of this preexisting system. A conversation is consequently an exchange of ideas mediated by signs (cf. de Saussure, 1966, p. 11 and figure 3.2).

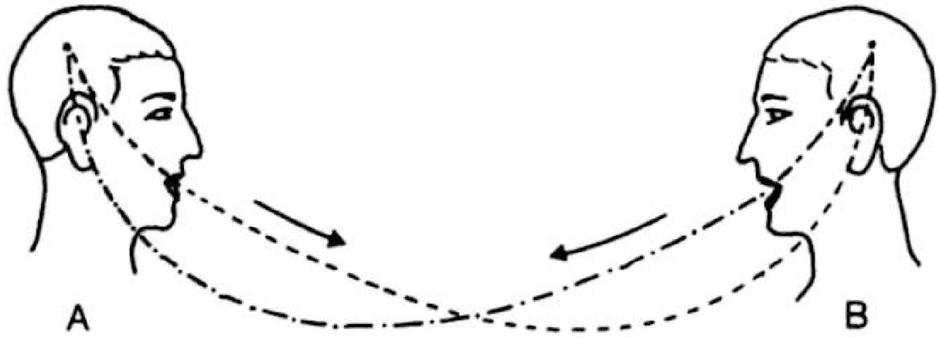


Figure 3.2: Saussure's picture of a conversation (taken from de Saussure, 1986, p. 11). He calls this ping-pong interaction the "speaking-circuit" and imagines it as a repeated encoding and decoding of "ideas" (de Saussure, 1966, p. 11–14).

- Language emerges from a universal human "language faculty": "[W]hat is natural to mankind is not oral speech but the faculty of constructing a language" (de Saussure, 1966, p. 10).

Not all of these assumptions were equally central to linguistics and philosophy, but by and large, they were the common ground under all theory of language. So to sum up, language was seen as a complex, opaque system confronting the individual. A person was assumed to have a linguistic loading dock which packed and unpacked thoughts. Communication, then, would be much like sending and receiving messages in Morse code.

3.2 Pragmatism

In this section, I introduce the philosophy of Ludwig Wittgenstein from the later part of his career, ranging from roughly 1939 to 1951. This part of his philosophy represents a radical departure from his early thought and, in a sense, from all of analytical philosophy.

3.2.1 Where to look for meaning

Let me begin with a simple fact: Some words can change meaning according to their context. This can be a shift between words senses (*I sat in the sun* versus *I looked at the sun*), but it can also be within the same dictionary sense: If you touch a scalpel with your fingertip, it may qualify as a *dirty scalpel*, but if you touch a pair of socks, they do not become *dirty socks*. This holds even though the word *dirty* is used in the same sense in both cases. Similar examples can of course be given with *polite*, *solid*, *genuine*, etc.

We can think about this phenomenon in many ways, but one option is to say that *dirty* can function in different domains, each having its own boundary between *dirty* and non-*dirty* cases. Thus, *scalpel*, *syringe*, etc. would belong to one domain, and *socks*, *shirt*, etc. to another.

But once we start thinking about such domains, more and more dimensions of difference appear. The word *mother*, for instance, will have radically different properties in different contexts: On one hand, it can be placed among words such as *mom*, *mommy*, *ma*, *mama*, etc., and then it will figure as one of the more formal and stiff alternatives. If we place it among words like *aunt*, *cousin*, *sister*, *grandmother*, etc., it does not have that property. Words like *offspring* or *stepmother* would provide further sorts of backgrounds that would again make the word appear differently.

At first glance, it seems that these families of associate terms single out ways to think of a *mother*. But this is one of the first misconceptions that we find Wittgenstein warning us about in his later work. Much of the *Blue Book*—his first attempt at a formulation of his later philosophy—is spend warning us about such "cognitive" ways of thinking about language:

> It seems that there are *certain definite* mental processes bound up with the working of language, processes through which alone language can function. I mean the processes of understanding and meaning. ... We are tempted to think that the action of language consists of two parts; an inorganic part, the handling of signs, and an organic part, which we may call understanding these signs, meaning them, interpreting them, thinking. (Wittgenstein, 1958b, p. 3; emphasis in original)

We find the same warning towards the end of his last book, *On Certainty*:

> There is always the danger of wanting to find an expression's meaning be contemplating the expression itself, and the frame of mind in which one uses it, instead of always thinking of the practice. That is why one repeats the expression to oneself so often, because it is as if one must see what one is looking for in the expression and in the feeling it gives one. (Wittgenstein, 1969, §601)

The thoughts or feelings that a word evokes are thus really just a trace of what we should be interested in, namely the *practice* of language. Most of Wittgenstein's later philosophy consists of reexaminations of traditional philosophical questions, which he analyses by relating them to linguistic practice. How these analyses are done will appear later in this section.

The most important tool Wittgenstein applies in his discussions of language is the concept of a *language-game*. Language-games are limited micro-languages he describes in order to shed light on the dynamics of actual linguistic practice. Mostly, he speaks about these games as complete, but primitive languages (cf. for instance Wittgenstein, 1958b, p. 81 and Wittgenstein, 1958a, §18). At other times, he seems to consider language-games as subparts of our actual languages, in particular fairly self-contained and isolated subparts (cf. for instance Wittgenstein, 1958b, p. 49 and Wittgenstein, 1958a, §23).

In either case, the point is to elucidate the problematic concept of "meaning" by showing how some kind of linguistic activity fits into a frame of practical activity. The simplest possible example is the straightforward situation that is described in the beginning of *Philosophical Investigations*:

> Now think of the following use of language: I send someone shopping. I give him a slip marked "five red apples". He takes the slip to the shopkeeper, who opens the drawer marked "apples"; then he looks for the word "red" in a table and finds a colour sample opposite it; then he says the series of cardinal numbers—I assume that he knows them by heart—up to the word "five" and for each number he takes an apple of the same colour as the sample out of the drawer. (Wittgenstein, 1958a, §1)

In spite of its simplicity, this example brings out the fact that at least one type of linguistic behaviour can be described with no reference to the classical concept of meaning. The people involved in the little scene need not worry whether their inner thoughts are really properly encoded in the words on the slip of paper, or whether they *really* understand the instruction—they just refine some non-verbal behaviour slightly by adding a little verbal behaviour.

Wittgenstein's works contain many more language-games defined in this top-down fashion. Some of them deal with fairly elaborate systems of orders and reactions, like the construction site games that are discussed in the *Brown Book* and *Philosophical Investigations*. Others deal with utterances that interact practice in more indirect ways, like the game in the *Brown Book* in which the players have two different ways of stating that they have done something, one for reporting the fact and one for establishing their capability.

3.2.2 The dependence of language upon life

One of the attractive features about the concept of the language-game is that it makes the relation between language and practical life very clear. Since the moves in a language-game have to be framed within a system of social interaction, the "meaning" of an expression cannot be chosen arbitrarily, but has to fill in a role which, so to speak, is left open by the practical concerns of the players.

So, for instance, the two people on the construction site described in the early paragraphs of *Philosophical Investigations* cannot invent some arbitrary distinction between $brick_1$ and $brick_2$ (on the basis of, say, the smell of the brick). Since this distinction does not relate to any practical concern in their life, it would not be intelligible to them.

Let me give a more familiar example. In English, an indirect object has to be capable of owning something, specifically the end product of the process that the verb describes. We thus see the following pattern of judgements (cf. Langacker, 2002, p. 14):

(3.1) (a) I sent the zoo a walrus.
(b) *I sent Antarctica a walrus.

(3.2) (a) I baked the couple a cake.
(b) *I baked the wedding a cake.

(3.3) (a) I wrote him a letter.
(b) *I wrote his address a letter.

We could be tempted to think of the difference between, say, zoos and continents as a "cognitive" or "grammatical" one, but Wittgenstein urges us not to: We do not reject the sentence *I sent Antarctica a walrus* because of its linguistic features, but because we do not know what it would mean in practice that a continent owned something. We would not know how to adjust our behaviour if Antarctica did or did not own a walrus.

This is emphasised if we imagine that the reverse judgements were actually made:

(3.4) (a) *I sent the zoo a walrus.
(b) I sent Antarctica a walrus.

(3.5) (a) *I baked the couple a cake.

(b) I baked the wedding a cake.

(3.6) (a) *I wrote him a letter.

(b) I wrote his address a letter.

This would be a strange language, but by no means incoherent or impossible. We could easily adapt to this new linguistic practice without changing much in our practical life—some words would have shuffled around. The fact that it feels much more natural to change our conception of the words than our practices also shows us how stable and entrenched these practices actually are.

We should also note the fact that the distinctions we make in our behaviour is much finer than the ones we observe in language. We should therefore resist the temptation to pigeonhole practice according to distinctions that immediately suggest themselves in language.

For instance, we suppose that the language-games available in life are already named by speech-verbs such as *assert*, *order*, *ask*, and *exclaim*, or indirectly enumerated by auxiliary verbs such as *be*, *must*, *can*, and *will*. But Wittgenstein tries to dampen our urge to build systems by hinting at the open-endedness of our practices:

> How many kinds of sentence are there? Say assertion, question, and command?—there are *countless* kinds: countless different kinds of use of that we call “symbols”, “words”, “sentences”. . . . Review the multiplicity of language-games in the following examples, and in others:
>
> Giving orders, and obeying them—
>
> Describing the appearance of an object, or giving its measurements—
>
> Constructing an object from a description (a drawing)—
>
> Reporting an event—
>
> Speculating about an event—
>
> . . .
>
> Translating from one language into another—
>
> Asking, thanking, cursing, greeting, praying. (Wittgenstein, 1958a, §23)

As theorists, we may frown at such a disorderly and incomplete list, but that is exactly Wittgenstein’s point: If we want to give a systematic and complete account of language, we will have to pretend that life is not life, but some restricted toy version of it.

A consequence is that it takes knowledge of situations to know when two utterances do the same. In a formal system, we may take pairs such as the following as equivalent:

(3.7) (a) Please pass me the salt.

(b) Can you pass me the salt?

(3.8) (a) I promise you that I will bake the cake for your birthday.

(b) I will bake the cake for your birthday.

But the relation between a sentence-form and its possible uses is not as clear as such well-behaved examples suggest. Try, for instance, to imagine situations where both of the sentences in the following examples would be natural:

(3.9) (a) Please marry me.

(b) Can you marry me?

(3.10) (a) I promise you that I will die before I turn 60.

(b) I will die before I turn 60.

We can of course brush aside such examples with a remark about "contextual effects," but that would miss the entire point: Real-life situations are not disruptions of an otherwise perfect language-system, but rather the fuel that drives it. "Only in the practice of a language can a word have meaning," Wittgenstein puts it (Wittgenstein, 1978, Part VI, §41).

The problems go deeper, too, as pointed out by an example by the Wittgenstein scholars Gordon Baker and Peter Hacker: Someone drills a hole in the ground, puts his arm into it, and says one of the following:

(3.11) (a) I feel water three feet down.

(b) I feel water fifty feet down.

In classical theory, the second assertion is perfectly meaningful, but happens to be false. That seems to belie our intuition that the sentence is not "untrue," but utterly bizarre, not any more meaningful than *Furiously sleep ideas green colorless* or *Good morning!* uttered in the same situation.

This can, of course, only be explained by reference to the situation: Understanding does not fail in the second case because some semantic or pragmatic module in the brain misfires, but because the situation is strange—we have perfectly good tools for coping with (3.11 a), but none whatsoever for coping with (3.11 b). It is meaningless in the real, phenomenal sense.

Faced with such examples, we can of course react by building even more theory: A more fine-grained taxonomy of speech-acts, a more detailed model of the context. But what Wittgenstein wants to hint at is not that our theories so far have been wrong, but that social life, in spite of its relative stability, does not work according to a set of rules, and we should not expect language to do so, either. This point will be elaborated more in section 3.3.

3.2.3 Utterance meaning: It does what it does

There is no clear dividing line between semantics and pragmatics in Wittgenstein's philosophy. Since understanding an utterance comes down to placing it in a frame of practical activity, the "meaning" of an expression is a sum of practical functions. In Robert Brandom's words, "appeal to *meaning* is just *one*, optional, theoretically laden way of dealing with *use*" (Brandom, 2008, lecture 2, p. 1; his emphasis).

That is, there is no such thing as meaning independent of practice. If two expressions have no functions, they can neither be functionally equivalent nor functionally different. This also has the consequence that understanding a sentence out of context becomes a matter of reconstructing its possible applications.

An sample analysis will bring this out more clearly. Consider the sentence

(3.12) If it rains, it rains.

An objectivist analysis of this sentence would reduce it to a logical tautology, so any effect it might have would not follow directly from the literal meaning. In contrast, in the Wittgensteinian analysis, the meaning of the sentence is equated with precisely its "pragmatic effects" in different language-games. In order to analyse the sentence, we should consequently look for the patterns of social verbal behaviour it fits into.

Note that there is a family of utterances in conditional form that work like promises in some respects and like predictions in other:

(3.13) (a) If it rains, I can get the tickets refunded.

(b) If it rains, I'm not going anywhere.

(c) If it rains, I'll drive you.

In spite of their superficial similarity, these linguistic moves are not moves in the same language-games. The conditions necessary to cancel them, for instance, are not the same:

(3.14) (a) It rains, but I cannot get the tickets refunded.
$\Longrightarrow$ The information I had does not apply (anymore).
$\not\Longrightarrow$ My initial prioritisation does not apply anymore.
$\not\Longrightarrow$ I am failing to live up to an obligation.

(b) It rains, but I am going somewhere.
$\not\Longrightarrow$ The information I had does not apply (anymore).
$\Longrightarrow$ My initial prioritisation does not apply anymore.
$\not\Longrightarrow$ I am failing to live up to an obligation.

(c) It rains, but I will not drive you.
$\not\Longrightarrow$ The information I had does not apply (anymore).
$\not\Longrightarrow$ My initial prioritisation does not apply anymore.
$\Longrightarrow$ I am failing to live up to an obligation.

If it rains, it rains thus fits into several different paradigms of behaviour. These can roughly be divided into two types, according to whether the moves in the game mainly

1. relate information to information, as in *If it rains* and *I can get the tickets refunded*, or

2. relate the information to an act somewhere between stating a fact and making a promise-like commitment, as in *If it rains, I'll drive you*.

If it rains, it rains contains no new information and would hence not be a felicitous move in a game of the first type. These games do accordingly not contribute to the meaning of the expression, and we will have to reconstruct the meaning from the games of the second type.

In order to understand the consequent *it rains* properly, we need to notice that situations can be construed more or less as a result of volition:

(3.15) (a) You're cheating!

(b) You sold me a car that doesn't work.

(c) That kid is getting into a lot of fights these days.

(d) Dogs are just like that.

(e) It appears that it's true.

Holding someone responsible for situations like these becomes more and more strange as we go from (3.15 a) to (3.15 e). This is, of course, not a distinction in language, but in life. In a culture where it was thought that people could cause assertions to appear true or false, there would be nothing strange about promising someone that *grass will appear to be red*.

For our purposes, however, the important point is that *it rains* is clearly perceived as being at the bottom of this scale: The weather is not seen as anyone's fault, merit, or responsibility. The utterance *it rains* can consequently not be used as a move of the same kind as *I'll drive you* or *I'm not going anywhere*. It does not put the speaker, the hearer, or anyone else under any particular obligation other than vouching for the fact.

But when we combine *it rains* with the condition *If it rains*, we have to arrive at an utterance that initiates some kind of social expectation, as argued above. Thus, we are forced to read the sentence as a kind of null expectation: The effect of the sentence is that in the case of rain, obligations, duties, promises etc. are left unchanged. Depending on which language-game the speaker is engaged in, the expectations that are preserved may be quite different:

(3.16) A: Your shift starts at nine o'clock tomorrow.
B: But what if the weather gets worse?
A: If it rains, it rains.

(3.17) A: I'll go for a hike tomorrow morning.
B: But what if the weather gets worse?
A: If it rains, it rains.

In (3.16), expectation is quite clear: B is obliged to show up at work regardless of the weather. (3.17) can be seen different ways, depending on why the hiker wants to inform the hearer about her plans. If the game is more like making a promise, then the utterance conveys that rain will not affect the promise. If the game is more like a self-description, then *If it rains, it rains* conveys that her dedication or energy is not affected by the rain.

This reading corresponds quite well to my intuitive understanding of the sentence as well as the examples I have found in corpus material and on the internet. One example occurs in the Michigan Corpus of American Spoken English (`http://quod.lib.umich.edu/m/micase/`). A construction of the form *P when P* comes up during a conversation about a computer programme designed to simulate a kind of auction:

> S1: (well but wait) the point is it's event driven. the auction [S3: okay] from the uh the auction is completely event driven. it receives bid when it receives bid it's uh does something [S3: right right right] and then, when quiescence, is reached it exits that, event driven mode and just does something. (MICASE, transcript MTG270SG049)

Here, as above, the point is that the programme does not solicit bids out of its own accord, as one might expect, but merely *receives bid when it receives bid*. It is a null condition parallel to the null expectation in the example *If it rains, it rains*.

3.2.4 The production and consumption of new meaning

If all utterances were standard moves in standard language-games, the difference between the objectivist and the Wittgensteinian philosophy of language would be cosmetic—explaining practice in terms of meaning or meaning in terms of practice would be equivalent. In reality, though, unfamiliar situations occur, and in familiar situations, people utter unfamiliar things.

This has some ramifications for the stability of our use of linguistic forms. In the *Blue Book*, Wittgenstein says:

> There is not one definite class features which characterise all cases of wishing (at least not as the word is commonly used). If on the other hand you wish to give a definition of wishing, i.e., to draw a sharp boundary, then you are free to draw it as you like; and this boundary will never entirely coincide with the actual usage, as this usage has no sharp boundary. (Wittgenstein, 1958b, p. 19)

The problem with defining *wishing* is thus not that we are bad at describing how we use words, but that these uses do not have sharp edges. Concepts surely come with a history of use, but in every new application, a real human being has to look at the context and decide whether the use is warranted.

To describe this phenomenon closer, Wittgenstein invoke two suggestive metaphors:

> ... games form a family the members of which have family likenesses. Some of them have the same nose, others have the same eyebrows and others again the same way of walking; and these likenesses overlap. (Wittgenstein, 1958b, p. 17)
>
> Think of the tools in a tool box: there is a hammer, pliers, a saw, a screw-driver, a rule, a glue-pot, glue, nails, and screws.—The functions of words are as diverse as the functions of these objects. (And in both cases, there are similarities.) (Wittgenstein, 1958a, §11)

Tokens of the same word do not perform the same function in all language-games, just as a fork does not do the same thing when it pierces a lobster and when it cleans your nails. There are, of course, family likenesses between eating a lobster, stirring your coffee, fishing a fly out of your soup, and cleaning your nails. But there are also differences, and a fork could not play the exact same part in all of these activities. We should see the different uses of the word *dirty* discussed in section 3.2.1 in this way.

These perceived likenesses are important, because language-games are not clearly delineated. Thanking or greeting someone is not an activity with a limited set of possible moves—alien tools may be imported. For instance, a greeting like

(3.18) Good early night!

is abnormal, but not unintelligible—especially if said by someone going to bed very early.

Wittgensteinian theory accounts for the strangeness by seeing greetings like *good night* as moves in a fairly self-sufficient game. It accounts for the intelligibility by the fact that we know plenty of other contexts in which *early night* or even *good early night* would be normal. Because we have a preunderstanding of the social situation, we are able to tell similar contexts from dissimilar.

That this last competence is necessary is evident from sentences like

(3.19) Good clear night!

Good clear night could certainly occur in other contexts, but as a greeting, it is hardly intelligible. It would be much harder to understand a person inserting *clear* than *early* into a greeting, and we would probably react more insecure and reluctant to *Good clear night!* than to *Good early night!*.

In general, the openness of language-games like greeting implies that there is no clearly defined set of uses for an expression:

> ... in general, we do not use language according to strict rules—it hasn't been taught to us by means of strict rules either. ... We are unable to clearly circumscribe the concepts we use; not because we don't know their real definition, but because there is no real 'definition' to them. To suppose that there *must* be would be like supposing that whenever children play with a ball, they play a game according to strict rules. (Wittgenstein, 1958b, p. 25)

In sum, we can think of utterances in their contexts as more or less standard ways of dealing with more or less standard situations. When we are exposed to nothing abnormal, nothing is ambiguous in the phenomenal sense. As the fit between situation and action becomes more odd, however, the utterances will become more ambiguous, and at some point, they become unintelligible. If we go even further, the action–context pairs will not even be recognised as utterances at all.

In this respect, there is no clear line between linguistic convention and convention in general. The former Danish supreme court justice Henrik Zahle has described a case in which he had to determine whether a manure tank was a *building* or not, because the construction of *buildings* above a certain height needed approval from the authorities. On one hand, there were plenty of arguments for the fact that the manure tank did not count as a building: It had no doors or windows, it could not be used for living, etc. One the other hand, it was very visible in the landscape, had traffic going to and from it, etc. In order to arrive at an adjudication, Zahle had to decide, rather than see, which factors were to count as essential (Zahle, 2003).

Hubert Dreyfus and Charles Spinosa provide another example, aksing

> ... is an instance of a door still an instance of that type when Hubert Dreyfus puts it on top of crates and starts using it as his desk? Our intuition that the desktop is no longer a door is only relatively stronger than our intuition that it remains a door, and we could start weighting more and more details that would strengthen either intuition. (Spinosa and Dreyfus, 1996, p. 743)

The question is open, not because we lack acces to the essence of doorness, but because the concept *door* occurs in an unfamiliar situation.

Both of these examples point to the fact that everyday language is optimised for everyday life. The fit between our linguistic resources and our concerns is not a coincidence, but a matter of cultural evolution. Accordingly, Wittgenstein tells us that "ordinary language is alright" (Wittgenstein, 1958b, p. 28). As he says, the expressions of our everyday language "are after all performing their office" (Wittgenstein, 1958a, §402).

3.2.5 Forms of life and the basis of understanding

In a discussion of nouns, Ronald Langacker notes (Langacker, 2002, p. 73) that count nouns like *cat* or *blade* may sometimes function as mass nouns. If the situation warrants looking at the cat as a more or less uniform mass, we do not have *a cat*, but *some cat*:

(3.20) After I ran over the cat with our car, there was cat all over the driveway.

More interesting, though, is his observation that such a shift in perspective might be naturalised and become permanent. To illustrate this, he imagines one termite saying to its termite friend:

(3.21) I don't like shelf – I'd rather eat table.

If this example seem too far-fetched, then consider these contrasts:

(3.22) (a) He bought a lot of chicken/lamb/fish.

(b) *He bought a lot of cat/dog/goldfish.

It is not hard to see that it would only take a slight change of eating habits for the ungrammatical sentences to become grammatical. The standard way of relating to your pet dog, however, is not to look at it as a piece of flesh or a quantity of "pet."

Eating dogs might be out of the ordinary, but the practice is still intelligible. The situation worsens if we encounter people with radically different practices that are not so easily translatable. In the *Remarks on the Foundations of Mathematics*, Wittgenstein imagines a group of wood-sellers that determine the price of a pile of wood by measuring its size. But then,

> ... what if they piled the timber in heaps of arbitrary, varying height and then sold it at a price proportionate to the area covered by the piles?
>
> And what if they even justified this with the words: "Of course, if you buy more timber, you must pay more"? (Wittgenstein, 1978, Part I, §149)

Such a difference in practice would make communication virtually impossible:

> How could I shew them that—as I should say—you don't really buy more wood if you buy a pile covering a bigger area?—I should, for instance, take a pile which was small by their ideas and, by laying the logs around, change it into a 'big' one. This *might* convince them—but perhaps they would say: "Yes, now it's a *lot* of wood and costs more"—and that would be the end of the matter. (Wittgenstein, 1978, Part I, §150)

This would probably be a very frustrating experience—but the issue is not, in Richard Rorty's words, a problem of "digging a little nugget of sense out of the mind" (Rorty, 1984, p. 55). The problem is rather to understand how it makes sense to live a life like these wood-sellers do. Their conception of *big pile* and *small pile* is unintelligible because we have no "agreement in action" (Wittgenstein, 1978, Part VI, §39).[1]

Wittgenstein's word for these cultural constraints is the German word *Lebensform*, which can be translated as both "life-form, organism" and "form of life, way of life." His thesis, then, is that we sometimes need to frame an utterance in an entire form of life in order to understand it: "the *speaking* of a language is part of an activity, or of a form of life." (Wittgenstein, 1958a, §23, emphasis in original; cf. also 19, 23, and 241, and pp. 174 and 226)[2]

This of course has consequences for his views on translation:

---

[1] While the story of the wood-sellers is a hypothetical example, Wittgenstein's contemporary, the British anthropologist Edward Evans-Pritchard, has described an actual experience similar to it (Evans-Pritchard, 1937, pp. 24–25). During his time with the Zande people in Sudan, he learned that if man was believed to be a witch, his father was, too, and *vice versa*. Hence, Evans-Pritchard inferred, if a male bloodline contains a single witch, the entire bloodline is automatically witches.

His Azande informants agreed with this line of reasoning, but still denied its conclusion—and did not see any problem in that. After some frustrating debates, Evans-Pritchard concluded:

> Azande do not perceive the contradiction as we perceive it because they have no theoretical interest in the subject, and those situations in which they express their beliefs in witchcraft do not force the problem upon them. (Evans-Pritchard, 1937, p. 25)

For some examples of Wittgenstein discussing actual anthropology, cf. his *Remarks on Frazer's Golden Bough* (Wittgenstein, 1979).

[2] Wittgenstein's notion of "forms of life" is notoriously controversial. Some philosophers, such as Richard Rorty, would agree with the interpretation I offer here. Others, like Nelson Goodman, would disagree.

> Now what characterizes an order as such, or a description as such, or a question as such, etc., is—as we have said—the role which the utterance plays in the whole practice of the language. That is to say, whether a word of the language of [a previously discussed] tribe is rightly translated into a word of the English language depends on the role this word plays in the whole life of the tribe; the occasions on which it is used, the expressions of emotion by which it is usually accompanied, the ideas which it generally awakens or which prompt its saying, etc., etc. (Wittgenstein, 1958b, §48)

If there is a similar expression in the target language, this is usually not a problem. But even then, serious problems may arise when the expression is used in abnormal way. In such cases, we will have to rely on our understanding of the social situation to pick a good translation.

This leaves us with a couple of questions: (1) What exactly does this preunderstanding consist in, and can we capture it in a theoretical model? (2) How often do we need this preunderstanding when reading a standard text, say a newspiece or a computer manual?

In the following two sections, I will address these questions one at a time. The first one is a quite philosophical question, and I will need to introduce some more theory to answer it. The second is a more empirical question about the mechanism of language, and I will answer it by giving some examples of linguistic phenomena that are directly influenced by our experience from real-life situations.

3.3 Existentialism and phenomenology

The philosophy of Ludwig Wittgenstein is a great tool for understanding how language builds on everyday life. In the next chapter, I will give some concrete examples of stabilities and instabilities in everyday life which surface as stabilities and instabilities in language.

Wittgenstein do, however, not have much to say about the concerns of practical life. For that, I will need to turn Martin Heidegger's philosophy, as expressed in his book *Sein und Zeit*. This will allow me to make the case that life, language, and cognition are neither isolated phenomena nor subjects that can be grasped in their entirety.

According to the introduction of *Sein und Zeit*, the purpose of the book is to ask for the meaning of "being." That is, not just why the sky "is" blue, why I "am" happy, or anything "is" anything, but what "being" means at all (Heidegger, 1973, p. 4). Heidegger wants to answer this question through a careful description of how people relate to the world around them.

This description is supposed to be faithful to the phenomena, but also prescientific and immediate: A talking person should be seen as a talking person, not a pile of organs emitting sounds. If we find ourselves talking in such physical or cognitive terms, we have already moved to far away from the phenomena, Heidegger warns (Heidegger, 1973, p. 16). The everyday world should be grasped, but not grasped to hard.

The inspiration for this style of investigation came from Husserl's phenomenology (cf. for instance Husserl, 1969). As we will see, phenomenology is not a type of introspection, though. Heidegger is not interested in "subjective experience," but in simpler phenomena, like people opening doors or walking on sidewalks. His attention to these aspects of everyday life results in s series of observations about the special human way of being. In tentative form, these are:

- human beings employ the things in the world as equipment;
- they intuitively perceive the equipment as related in a certain way;

- their perception of meaning of a piece of equipment comes from its role in this network;
- every network of equipment corresponds to an existential role;
- these existential roles are unstable.

In the following subsections, I will explain these points in more detail and by concrete examples.

3.3.1 The interactive environment

The first phenomenological observation Heidegger does is that we mostly relate to things around us as things *for* something. The natural way to relate to a hammer is to pick it up and hammer with it, not to look at it as metal blob on a stick:

> The hammering does not simply have knowledge about the hammer's character as equipment, but it has appropriated this equipment in a way which could not possibly be more suitable. ... [T]he less we just stare at the hammer-thing, and the more we seize hold of it and use it, the more primordial does our relationship to it become, and the more unveiledly is it encountered as that which it is—equipment. (Heidegger, 1973, p. 69)[3]

The possibility of just relating to the hammer as a substance with some properties is available, but it takes an effort: When you opened the door to the room you are in, you did not "decide" to use the latch (cf. p. 67). The everyday dealing with doors, handles, and hammers has "*its own* sight" and "its own 'knowledge'," as Heidegger says (p. 69 and 67).

Heidegger introduces the term *ready-to-hand* for the way equipment appears to us. When things appear as objects with "substantiality, materiality, extension," etc., calls them *present-at-hand* (p. 68). This pair of concepts is roughly equivalent to the ones that began appearing in English-language philosophy some twenty years later, know-how/know-that and tacit/verbal knowledge (Ryle, 1949, ch. 2; Polanyi, 1966).

As human beings, we are thus placed in a kind of interactive environment. This environment is populated by things that we immediately understand as equipment.

The concept of equipment should taken in a very broad sense, including rooms as "living-equipment" and streets as "equipment for walking" (Heidegger, 1973, pp. 68 and 107). The interactive environment is consequently not just found in "the homely world of the workshop, but in the *public world*"—streetlights serve a purpose because we use light, roofs because the need protection from the weather, factories because we use commodities, and so on (p. 71).

Such ready-to-hand things appear differently when they function as streetlights, roofs, and factories than when they function as, say, physical systems of bodies with masses. This has consequences for sound, space, colour, and the like:

[3] *Sein und Zeit* is full of puns and neologisms, and translating it is a notorious pain. I have used the standard translation by John Macquarrie and Edward Robinson (Heidegger, 1973) as reference, but felt free to change the translation when I thought it missed the point. I have tried to stick consistently to the following glossary: *beings* = *Seienden*; *equipment* = *Zeug*; *primordial* = *ursprünglich*; *encounter* = *begegnen*; *first/firstly* = *zunächst*; *environment* = *Umwelt*; *ready-to-hand* = *zuhanden*; *present-at-hand* = *vorhanden*; *totality of equipment* = *Zeugganzheit*; *referential totality* = *Verweisungsganzheit*; *its being is at stake* = *es geht um sein Sein*; *interpretation* = *Auslegung*; *seeps into the possibilities* = *dringt in die Möglichkeiten*; *always already* = *immer schon*. The page numbers refer to the German version.

> What we ‘first’ hear is never noises or complexes of sounds, but the creaking waggon, the motorcycle. One hears the colomn on the march, the north wind, the woodpecker tapping, the fire crackling. ... Even [when we hear speech which] is indistinct or in a foreign language, what we first hear is unintelligible words, and not a multiplicity of sound data. (pp. 163–64)

> To the person who uses glasses, for instance, which are in terms of distance so close that they “right in front of his nose”, this piece of equipment is environmentally speaking farther away than the picture on the wall across the room. (p. 107)

The same goes for, for instance, the phone in your hand or the street below your feet: To one stance, they are evidently right next to you, but to another, they are not even there. Only when the phone stops performing its function, it becomes visible again (pp. 107 and 75).

This also shows how the present-at-hand might be derived from the ready-to-hand. Equipment which is useless, unavailable, or in the way may disturb an activity without halting it. The activity can thus still provide a pragmatic frame of reference allowing the equipment to make sense even though it appears as unfit (cf. p. 73). A broken key appears as a broken key, but a pulverised key appears as dust.[4] Given enough situations with enough different pragmatic concerns, we may learn to apprehend more and more properties of present-at-hand keys.

In classical philosophy, this would be explained as a case of “seeing the same thing, but interpreting it differently.” But it is noteworthy that Heidegger does not assume any kind of “seeing” apart from the familiar one that sees a hammer as a hammer. Physically speaking, we do not receive anything but energy and chemical compounds from the world—neither “sensations” or “experiences.” When we assume that there is some level of uninterpreted sense data that we have direct access to, that is either in a degenerate sense with very little content or an unwarranted assumption as plausible as the claim that we can see the inside of our own eyeballs (cf. Dreyfus, 2007, sec. VIII; Merleau-Ponty, 1962, pp. 3–74).

In Heidegger and Husserl’s thought, on the other hand, abstract properties like spatial extension grow out of our preunderstanding of the ready-to-hand (Heidegger, 1973, p. 71). In order to understand geometry, for instance, you need some intuitive familiarity with real objects like strings, rulers, boxes, and sheets of paper.

That should occasion some scepticism about the project of collecting all the facts in the world: If facts about things are really aspects of practical situations, then no amount of staring at a bike will explain why I can ride it without a saddle but not without pedals. In the next subsection, I will elaborate this argument a little more and point out some implications of it with respect to language.

3.3.2 Purpose, reference, and holistic understanding

Equipment is always equipment *for* something. In Heidegger’s terminology, it always has an “in-order-to” (Heidegger, 1973, p. 68). Things can accordingly only function as equipment in a context of means and ends:

> Strictly speaking, *one* piece of equipment never “is.” To the being of equipment there always belongs a totality of equipment, in which this piece of equipment can be, in which it is. ... Equipment is with regard to its equipmentality always *because of* the

[4] Similarly, we might notice when the other team catches the ball in a game of basketball, but not necessarily if a gorilla walks through the field (Simons and Chabris, 1999).

> affiliation with other equipment: Writing equipment, feather, ink, paper, pad, table, lamp, furniture, window, doors, room. (p. 68)

This ontology is not a system of connections between present-at-hand things, but an "referential totality" between ready-to-hand pieces of equipment (p. 70). You can thus use a screwdriver in-order-to pull out a screw out in-order-to solder a cable in-order-to fix a microphone without thinking about the material properties of either.

This in-order-to also provides a context that makes human skills both flexible and stable. One one hand, we do not need an inference to tell us that we can pound a tent peg down with a rock instead of a hammer—on the other hand, nor do we need it to see that we could not do it with a handful of sand. Similarly, we would be more likely to win a soccer match if we killed the opposing team, but because of our intuitive understanding of the in-order-to of soccer, we do not even think in such terms.

This familiarity with the in-order-to's affects our relation to distance, solidity, brightness, etc. just like our understanding of the equipment did. Heidegger uses the east–west orientation of churches as an example: The fact that cemetery lies on the sunset side of the church "refers" to the trajectory of the sun (Heidegger, 1973, p. 104). But there are simpler examples: A table can be too high or too low, but only because it "refers" to chairs, chopping boards, books, and other pieces of equipment. If we always sat on the floor when we ate and worked, the referential totality and hence the natural "table-height" would be different.

All of these points radicalise Heidegger's holism. Latches, tables, bikes, buildings, streets, street-lights, and stars are what they are only relative to a totality of equipment in a totality of references. There is no intelligible, but neutral base level we could explicate everything in terms of. As Sartre puts it, we cannot "attain the essence by heaping up the accidents" (Sartre, 2002, p. 4). Any attempt to describe equipment as self-sufficient substances with context-independent properties would either have to describe everything in the world at once or smuggle background knowledge in through the back door.

Given Heidegger's perspective on human conduct, it is natural to see communication as a part of the absorbed, everyday dealings. Indeed, Heidegger notes that signs play a role like any other ready-to-hand equipment. The difference is that signs do not just refer to things the way scissors refer to paper, but make such references explicit (p. 80–82). But since the references are pragmatically defined and intuitively understood, this means that signs are, too. That has some profound consequences.

As an example, he takes a knot in a handkerchief (p. 81), a mnemonic that would be perfectly familiar to the reader in 1927. The reference or in-order-to of such a sign will turn out to be extremely vague if we try to put it into context-independent propositions, but because it is used within the frame of everyday dealings, it does not have to be spelled out. It may even be very precise in its own pragmatic sense.

We should see utterances the same way. A sentence like *The hammer is heavy* can be taken as a categorial judgement along the lines of "The hammer-thing has the property of heaviness," but Heidegger comments:

> In the careful circumspection there are no such assertions "at first". Yet, it does have its own type of interpretation, and in relation to the just mentioned "theoretical judgement", it could have sounded: "The hammer is too heavy!", or better even: "Too heavy!", "The other hammer!". (p. 157)

In such situations, the utterance offers itself as an auxiliary tool when other things become *un*ready-to-hand. There certainly is a "knowledge-system" behind an assertion, but it is knowledge of the ready-to-hand, not the present-at-hand.[5]

As we saw above, a metal blob on a stick only becomes a hammer when it hammers. We understand it only insofar that we see it as relating to real nails, real wood, and real houses. If not for the trace of this involved action, we could not understand the "hammer-thing" as a hammer.

Since sentences are tools, too, they are only understood sentences when they make relations in real life explicit. Any "meaning" they have is a trace of these real-life relations. Understanding, as Heidegger says in the chapter on the topic, always elaborates "*something as something*" (Heidegger, 1973, p. 149; his emphasis). This implies that understanding hides as much as it shows.

For instance, if we understand politics as a "game" or education as a "vaccination," we invoke the entire holistic concepts of games and vaccinations. That draws on all of our first-hand experience with games and vaccinations in real life. Any attempt to spell out this understanding would itself elaborate "something as something" and cover up a different part of this background knowledge (cf. Dreyfus, 1980, 8–10 and 15–17).

No amount of grammar or corpus data could therefore reconstruct our familiarity with language, just as no amount of piled up facts could make a hammer out of a blob on a stick. We learn language as participants in real situations saturated with risk, context, and significance. It is existential not linguistic experience that teaches us to see the huge difference *Is there any meat in this?* and *Is there any salt in this?*.

Counting how often expressions occur among other expressions depicts language as a present-at-hand object—a pile of expressions with "meanings"—instead of equipment in use. This will give us traces of utterances, but never get to the utterances themselves. In order to grasp language, life must be grasped, too. This has to happen all at once or not at all.

3.3.3 Cognition in the existential openness

As mentioned above, all equipment has an in-order-to reference, like the power socket to the fridge, the fridge to the milk, and the milk to the coffee. If we want to formalise all the know-how that goes into a mundane activity like changing a fuse, we need to explicate this structure in its entirety.

One of the reasons that this seems hard is that the in-order-to references are not always very clear or very definite. A can opener opens cans, and a screwdriver screws, but what exactly does a hat do? What is the in-order-to of a novel, a doll, a lecture, a wedding ring, or a spacecraft? We may of course shoot from the hip at each of these questions with a superficial biological or sociological explanation, but that still leaves the phenomenological issue to be dealt with. Something substantial may be going on in the normality of putting on clothes or reading the newspaper.

Heidegger notes:

> But the totality of involvements itself goes back ultimately to a towards-which in which it has *no* further involvement . . . The primary "towards-which" is for-the-sake-of-which.

[5] Put in more psychological terms, communication requires a "joint attentional frame" (Tomasello, 2003, pp. 21–26 and 65–67). A child's elliptical chattering is not a suppressed version of adult speech or of a full-fledged logical analysis. By definition, an attentional frame does just not include everything. This is not a shortcoming that has to be "fixed."

Heidegger's take on abstract thought generally resonates a lot with developmental psychology. Lev Vygotsky had some more sophisticated things to say about the difference between social and egocentric speech, but there is a large overlap in their ideas. Vygotsky's *Thought and Language* (Vygotsky, 1986) was published in the USSR seven years after *Sein und Zeit* came out in Germany.

> But the "for-the-sake-of" always pertains to the being of *Dasein* [Heidegger's term for a human being], for which, in its being, that very being is always *at stake* (Heidegger, 1973, p. 84).

In other words, Heidegger suggests that the ultimate in-order-to of all our activities is our own being, in the sense of "identity" or "existence." Or, to put it in slightly anachronistic terms, the equipment allows you to "perform" some role or identity (Butler, 1999, part three, ch. IV).[6]

Take for instance the role as citizen of a particular nation. Many of us have been brought up to loathe explicit signs of nationalism, such as military parades and thick political rhetoric. In certain situations, we nevertheless still intuitively behave as citizens of a nation, for instance if we are abroad and pass someone speaking our language, or see an article on our own country in a foreign newspaper.

This mechanism only works because we constantly encounter ready-to-hand equipment whose in-order-to's refer to the nation: The dictionary, the national currency, the stamps, the passport, the border patrol, the newspaper's distinction between national and international news, terms like *here* and *us* for the nation, etc., etc. (Billig, 1995). None of these are thematised or explicit—and if they were, they would most likely not work.

Hubert Dreyfus and Charles Spinoza gives another example:

> For instance, when we walk off a crowded street into a cathedral, our whole demeanor changes even if we are not alert to it. We relax in its cool darkness that solicits meditativeness. Our sense of what is loud and soft changes, and we quiet our conversation. In general, we manifest and become centered in whatever reverential practices remain in our post-Christian way of life (Dreyfus and Spinosa, 2003, p. 346).

In both cases, the equipment discretely invites us to perform a certain identity, and we can respond to the call or ignore it.

That is the sense in which our existence is "at stake" in the acts of going to work in the morning, brushing our teeth, or putting on clothes. We might have been the kind of being that figured that teeth get dirty anyway and that clothes are annoying—but we happen not to be. By the same token, it is not a quick cost–benefit analysis that makes a parent run into a burning house to save a child; and we would not know what to tell the father who stopped feeding his kids because he found out that they were going to die sometime anyway (cf. Dreyfus and Spinosa, 2003, p. 345). That is, in a very basic sense, simply not "what one does."

This connotes a certain sense of conformity that Heidegger is not blind to:

> We enjoy and amuse ourselves as *one* does; we read, look at, and judge literature and art the way *one* looks at and judges it; but we also withdraw from "big crowds" the way *one* withdraws; we find are "shocked" by that which shocks *one*. The one which is no one in particular and the everyone, although not the sum, prescribes the mode of being of everydayness. (Heidegger, 1973, p. 127)

Since everydayness does not just come with a set of habits, but with a set of prescriptions, too, the identities provided by it are natural and easy to assume. Other people will support you in "doing 'being normal'" and expect you to return the favour (Sacks, 1995, p. 215).

[6] My interpretation of the more existentialist layer in *Sein und Zeit* is influenced by Hubert Dreyfus, Charles Spinosa, and William Blattner, but it is far from a universally approved reading. There is plenty of evidence to support it, but sceptical readers can think of it as an application instead an interpretation of Heidegger's work.

The theme of negotiating your identity is one that Heidegger picked up from Søren Kierkegaard and Friedrich Nietzsche, who were both very sceptical about standard rational justifications of existential choices. With respect to the things that really matter—the ones that put your identity at stake—we choose a justification rather than justify a choice, they argued (Kierkegaard, 1978, pp. 65–112).

A couple can argue perfectly rationally about the assertion "You're not paying attention to me" without ever getting anywhere, for instance. In such cases, no amount of rigourous method will help you take a decision. Science, reason, art, love, or a burning bush might advise you, but you still choose to follow its advice yourself (cf. Sartre, 2007, p. 35).

This has consequences for cognition. Remember that the intelligibility of a thing is based on the entire totality of equipment: If chairs looked differently, we would relate differently to tables, and if there were no cans, can openers would make no sense at all. All of these internal references within the totality of equipment then turned out to be organised around the existential roles they allow us to perform.

Further, linguistic differences sometimes mirror existential differences, as Kevin Knight alludes to here:

> [A] Japanese-English machine-translation system I have worked on ... once translated a Japanese sentence as "there is a plan that a baby is happening in her"—a reasonable translation but with a definite Japanese-semantics feel to it. (Knight, 1997b, p. 83)

Not only are such existential roles varied, they are also under constant negotiation. In the Heideggerian jargon, the way of being "seeps into the possibilities"—it constantly, spontaneously proliferates new ways of relating to the world (Heidegger, 1973, p. 145). Because truth, relevance, and intelligibility all emerges from the way we relate to the world, new practices do not just fill out previously given slots, but let us see the world in a new way.

As a simple example, take a macaque monkey that sees a potato being washed for the first time (Kawai, 1965). On the basis of just that exemplary piece of behaviour, the monkey may change its behaviour and thus begin to see unwashed potatoes as "dirty" and water as a resource for washing as well as drinking. After such an existential shift, the network of references in the monkey's totality of equipment will have changed.

Since human existence is subject to this openness, so is our practice, cognition, and language. Enumerating all the facts of the world is not just a lot of work—it literally presupposes a set of rules for how to be a human being. Anything might develop new meaning when people relate to the world in new ways. In a concrete text or conversation, we might always be open and able to take on such new ways.

3.3.4 Hermeneutical footing

According to objectivist philosophy, an utterance is a container that "conveys meaning": We transmit thoughts by the means of speech. But that is as misleading as saying that a ball "conveys tennis."

Language should not be depicted encoding scheme. Not because it is too autonomous and misbehaved to contain our thoughts, but because thoughts are irrelevant to the discussion. The utterance *Too heavy!* is already "out there," relying on the structure of the ready-to-hand world. An analysis which attributes its meaning to anything else will watch it slip through its fingers.

If we really want to understand a conversation as a ball game, a game of squash would be a more appropriate image. Human beings are, to talk Heideggerian, "always already" in a situation

(Heidegger, 1973, p. 135ff). We bring a history and a set of future possibilities into any particular moment. We always have a hermeneutical footing in the world. These existential resources are put to work whenever we understand something: As soon as we recognise something as a text—even just as an unintelligible text—we already have an understanding of the situation.

All texts are thus always already in situations. A texts is just as much an encountered beings as a platypus or a slap in the face. Just as there is no way to look at a thing and see nothing at all, there is no way to read a text and see no meaning at all: "When I say that there is nothing outside the text, I mean that there is nothing outside the context, everything is determined," Derrida explains (Derrida, 1999, p. 79).

Examples such as the those on page 56 are thus completely natural. Meaning change with context. Of course. Meaning *is* context. Words, things, and people attain their meaning from the way they gear into situations, nothing else. If we understand anything at all, we understand it on the background of our hermeneutical footing in life.

To get a sense of what this resources, think of the options facing an anthropologist has when she want to familiarise herself with a new culture. She might ask her informants about their culture and collect as much statistical data as she could. That would give her some verbal accounts of how these people feel about weddings, funerals, and soccer matches, and possibly some quantitative facts. That would be sufficient if the anthropologist was always able to understand this information in terms of her own experience.

In a less orderly world, though, participant observation would be a better tool (DeWalt and DeWalt, 2002). Instead of asking for descriptions, the anthropologist could join her informants at weddings, funerals, and soccer matches and experience the significance of such events rather than listen to descriptions of it. She might suddenly come to realise what this strange tribe really meant when they threw around words like *commitment*, *God*, or *our team*.

We can approach a language the same way—in fact, the tradition of Franz Boas, Edward Sapir, and Bejamin Lee Whorf has already done so. If we want to understand language, we need to understand the everyday world it works for, even when it is our own. There is no way to isolate the experiences that will be relevant for understanding an utterance: Spicy food, right-handedness, pet animals, and bus travels are all a part of the shared background that language feeds on. Any such preunderstanding may mobilised to give meaning to a new use of a verb tense, a new subcategorisation scheme, a distinction between two genitive forms, an amalgram, and so on.

In the following chapter, I will give some concrete examples of how this interaction between life and language looks. In terms of linguistic theory, this will invite a perspective akin to that of cognitive grammar but counter the temptation to look at language as something "in the head." In a sense, it is of course our brains that make us stop at the red traffic light or use passive constructions, but if we really wanted to explain the fact, that would be a bad place to start.

Chapter 4

Examples of the life–language interface

The theories I discussed in the previous chapter depict language as a social tool specialised for the practical everyday life of a particular form of life.

The existential roles available in this form of life are shaped by two opposing forces: An expressive, creative, productive inclination in the individual, and a levelling, controlling, and normalising inclination in the social environment. The levelling tendency creates stability and intelligibility in life and thus also in language. The creative tendency creates openness to new existential possibilities and thus allows for production of new meaning.

In this chapter, I describe some linguistic mechanisms that produce new meaning in everyday conversation. These mechanisms give rise to lexicalised forms over the course of slow, historical changes, but they also proliferate more idiosyncratic and short-lived forms on the level of the individual sentence. They consequently pose a problem to all models of language that are cast in terms of past linguistic regularities rather than familiarity with everyday life.

4.1 Metaphor

In old textbooks on rhetoric or linguistics, metaphor is typically defined as a simile (*A is like B*) in which the comparison marker *like* has been dropped (Cody, 1905, p. 110; Opdycke, 1949, p. 419). That theory accounts for metaphors in the form of *Jesus is my compass* or *that car is a piece of shit*.

Aristotle's classical definition of metaphor is slightly more sophisticated. He thought of metaphor in geometric terms, as an equation between a relation $a : b$ and a relation $c : d$ (Aristotle, 1902, p. 41–42). For instance, take the following equality of relations:

(4.1) *evade* : *tax* = *dodge* : *projectile*

Using this equality, we can create the metaphor *Walmart dodges taxes* by performing the substitution

(4.2) Walmart evades taxes $\longrightarrow$ Walmart dodges taxes

Aristotle's theory permits us to talk about some additional forms of metaphor, but is based on the unanalysed notion of similarity just like the 20th century theories.

Such similarity-based theories have some appealing features, including intuitive plausibility and a fairly clear-cut distinction between metaphor and metonymy. The phrase *the light of reason*, for instance, plays on a similarity between light and the effects of rational thought, so it applies *light* metaphorically. The word *skylight*, on the other hand, does not—there is no similarity relation between a light and a window, only the relation that windows usually provide light. The word *skylight* is hence derived metonymically from the concept of *light*.

As I mentioned in the previous chapter, Heidegger argues in *Sein und Zeit* that understanding always is an elaboration of something "as" something. He further argues that competent use of language applies such understanding. If we put those two ideas together, we should expect that competent use of language elaborates things "as" other things.

Since metaphor may be seen as asserting that something is "like" something, we should expect to see metaphor everywhere people talk about the world as they understand it. This is indeed what George Lakoff and Mark Johnson report in their influential book *Metaphors We Live By* (Lakoff and Johnson, 1980, p. 3). For instance, they point out that when we talk about time, we constantly borrow concepts that relate to money:

(4.3) (a) How do you *spend* your time these days?

(b) That flat tire *cost* me an hour.

(c) I've *invested* a lot of time in her.

(d) I don't *have enough* time to *spare* for that.

(e) You're *running out* of time.

And so on. Evidently, this is a productive mechanism, not just a matter of a finite set of idioms: I can *waste*, *buy*, *borrow*, *value*, and *gamble with* my time, as well as indefinitely many other things. Lakoff and Johnson attribute this mechanism to an underlying non-verbal mapping that they label TIME IS MONEY.

The metaphors or the underlying mappings are not always obvious. We regularly *handle*, *tackle*, and *sort out* situations; we *run*, *execute*, or *kill* processes; and we *lean towards*, are *attracted to*, and *repelled by* options. Moreover, such metaphors a not an aesthetic frosting on the cake, but central to our expressive powers: There is hardly any way to explain what it means to *use your time* in literal terms. We do not invoke the concept of *using time* to explain a phenomenon different from *using time*.

According to Lakoff and Johnson, then, metaphor is not the peripheral, poetic mechanism that we might intuitively take it for. Metaphorical meaning production naturally occurs whenever you tell me how to change the batteries of a flashlight, how I can recognise your coat in a pile of similar coats, how acrylic paint differs from water colour, and the like. Metaphors are used when a tailored word is not around, and we reach for the nearest other piece of equipment.

In order to get an impression of roughly how pervasive the use of metaphor is, I have scanned the MICASE corpus, collecting 50 occurrences of the word *fight*, and then counted how many of these that were literal versus metaphorical.

I counted an instance as literal when it referred to *fights* in which it was an essential component that you would be physically beaten, stabbed, shot at, shoved around, etc. Examples include

(4.4) (a) that's a Largemouth Bass. it's gonna fight a lot harder (LAB175SU032)

(b) forty-one of them feature fight scenes. (COL140MX114)

(c) i hit him in the back of the head with an apple, in an apple fight (SGR200JU125)

(d) in the Civil War, um blacks they wanted to fight for their rights (STP095SU139)

The rest I counted as metaphorical. These included instances such as

(4.5) (a) you could fight over what what a good definition of a species is (OFC175JU145)

(b) you know more can be done to fight crime maybe the police need more money, right? (LES220SU140)

(c) an encounter between an m- a T-cell and an antigen, that could help fight off infection. (COL200MX133)

(d) we had to fight, not to have to publish. in order to get tenure you know (COL999MG053)

The tally showed 17 metaphorical uses, 32 literal, and 1 ambiguous (*she's not trying to fight um, she's not opposing, the power* in SEM300MU100).

The word *fight* is thus used metaphorically more than a third of the time. For rarer words like *television*, *country*, *judicial*, *discontent*, or *grandmother*, this ratio would almost certainly be smaller. But for words like *handle*, *topple*, *kill*, *trap*, *star*, *light*, *sweet*, *hard*, *build*, *health*, *follow*, and indefinitely many other words, I suspect that the ratios would be just as large or larger than it is for *fight*. If that conjecture is true, metaphor is indeed a very widespread phenomenon.

My examination of the occurrences of *fight* in the MICASE corpus is of course just a small probe investigation and should be taken with a grain of salt. However, even in this small sample, the *fight*-metaphors reveal a diversity of expression, covering topics such as crime, disease, ideas, stereotypes, bureaucracy, software bugs, intellectual arguments, and exam questions. These different applications are often quite qualitatively different. Fighting a virus and fighting crime is not the same thing, and often has different translations. *Fighting* in the context of war and in the context of a schoolyard fistfight would be expressed differently in Danish, for instance. The wide scope of a single metaphorical concept thus poses a practical problem for machine translation.

This productivity seems surprising when we look at an individual lexeme and marvel at its versatility. However, if we look at a speaker as a pragmatically engaged agent using whatever concepts she finds ready-to-hand, this versatility follows naturally from the diversity of everyday practice. For instance, how would you refer to the piece of plastic covering the batteries of your computer? A *cap*, a *lid*, a *hatch*, a *cover*? What would you call the part of a broom that touches the floor—the *head*, the *foot*, the *end*, the *brush*? Language itself does not dictate any particular choice, because language does not do anything at all—it is just an inventory of equipment that may or may not serve our purposes (cf. Rorty, 1993).

Thus, our first-hand experience of the world is constantly at work when we produce or consume texts like the following:

(4.6) This item provides protection against viruses that try to write to the boot sector and partition table of your hard disk drive. You need to disable this item when installing an operating system.

Concepts like *provide*, *protect*, *virus*, *write*, *sector*, *table*, *item*, *install*, *operate*, etc. all have familiar meanings in the everyday world of working, shopping, living, writing, and driving. When they are used in a computer manual, we need this familiarity to relate the word to its context in the right way.

To people familiar with diseases, pencils, blackboards, etc., this instability of expression is not a problem. But if we treat all meaning as old meaning, we soon run up against the problem that the everyday world is unanalysable and existentially open: Any formalisation of the lexical openness of concepts will itself be open.

A couple of examples of mistranslations of metaphors may serve to underscore this point. All of the following examples are German-to-English misinterpretations of metaphors done by Google Translate on December 21st, 2009.

(4.7) Lassen Sie mich jedoch sehr *deutlich festhalten*, daß dies... (= "hold clearly" = make clear)
However, let me *hold it very clear* that this...

(4.8) ...daß dies eine Sache ist, die die Bürger wirklich *berührt*. (= "touches" = concerns)
...that this is something that citizens really *touched*.

(4.9) Auch der Imageschaden wirkt *belastend* für ihn. (= "burdening" = incriminating)
The loss of image does *stressful* for him.

(4.10) Keine *Regelung* ist somit die schlechteste Lösung. (= "rule, system" = regulation)
No *scheme* is therefore the worst solution.

(4.11) Mitarbeitern werden eine böswillige Absicht *unterstellt*. ("under-placed" = suspected for)
Employees be *placed under* a malicious intent.

(4.12) Um das Betriebsklima zu *fördern*,... (= "promote, advocate" = nurture)
To *promote* the business climate,...

(4.13) Sie *holt* vom Mitarbeiter die Einwilligung ein. ("fetch, get, take" = obtain)
It *brings* the employee *receive* the prior agreement.

I have ignored the problems that are most likely not due to metaphorical processes.

4.2 Similarity

In what I have said so far, I tacitly accepted that metaphors should be described in terms of similarity. That theory has, however, been looked upon with suspicion by a number of scholars, including George Lakoff and Mark Johnson, who carefully avoid using it as an explanatory tool. The philosopher Nelson Goodman has even overtly dismissed similarity as "a pretender, an imposter, a quack ... professing powers it does not possess" (Goodman, 1972, p. 437).

One of the reasons that similarity is a problematic tool to rely on is that it is not stable. As the psychologist Douglas Medin notes,

> ...a *zebra* and a *barberpole* would be more similar than a *zebra* and a *horse* if the feature "striped" had sufficient weight. This would not necessarily be a problem if the weights were stable. (Medin, 1989, p. 1473)

But the problem is that they are not. In experiments, he and Edward Shoben found that people would change their similarity estimates wildly with context. *Black* and *blue shoes*, for instance, are much more similar than *blue* and *green shoes*, but the exact opposite holds for *black* and *blue eyes* versus *blue* and *green eyes*. Also, *grey* and *black clouds* are found to be more similar than *grey* and *white clouds*, while the opposite holds for *hair* (Medin and Shoben, 1988, pp. 170–75).

So contexts or situations change how we relate to things. This is, however, only surprising if one sees similarity through the optic of an essentialist theory where objects appear before us as self-sufficient substances with context-free properties. In reality, it takes an act of alienation to produce present-at-hand greyness out of ready-to-hand clouds and hair. Medin, too, is aware of this fact:

> It is perhaps only a modest exaggeration to say that similarity gets at the shadow rather than the substance of concepts. Something is needed to give concepts life, coherence, and meaning. (Medin, 1989, p. 1474)

This "something" is our experience of the everyday phenomenal world, including clothes, eyes, rain, and aging. One does not cut a steak with a pair of scissors or sing in a phone booth, but one does cut grass onion with scissors and sing in the shower; not because it is inherently meaningful, but just because our life-world happens to be structured that way.

Another problem with relying on similarity for explanation is that it does not actually explain anything. As Thomas Kuhn puts it, "anything is similar to, and also different from, anything else" (Kuhn, 1977, p. 307).

There is, for instance, some metaphorical truth to the sentence *your money is your servant*, but equally much to *your money is your master*. Appealing to the notion of similarity does not clarify why. Reversely, nor does it clarify why the concept of a *high-pitched voice* seems more natural than the concept of a **thin-pitched voice*.

A natural conclusion is that, as Goodman says, "[m]etaphorical use may serve to explain the similarity better than—or at least as well as—the similarity explains the metaphor" (Goodman, 1972, p. 440). Lakoff and Johnson make the same observation:

> It is important to remember that correlations are not similarities. Metaphors that are based on correlations in our experience define concepts in terms of which we perceive similarities. (Lakoff and Johnson, 1980, pp. 151–52)

This means that metaphor institutes similarity rather than *vice versa*. As an example, they consider the way we talk about ideas in terms of food: *half-baked ideas*, *fishy arguments*, *theories that you can sink your teeth into*, etc. (Lakoff and Johnson, 1980, pp. 46–47). Since there is nothing natural in the notion of smelling an argument or eating an idea, these verbal acts need to be ascribed a power of their own:

> Thus, the IDEAS ARE FOOD metaphor establishes similarities between ideas and food. Both can be digested, swallowed, devoured, and warmed over, and both can nourish you. These similarities do not exist independently of the metaphor. The concept of swallowing food is independent of the metaphor, but the concept of swallowing ideas arises only by virtue of the metaphor. (Lakoff and Johnson, 1980, pp. 147–48)

In other words, our non-verbal experience provides the resources, but it does not dictate the use. For instance, the concept of an electrical current *flowing* through a wire like water flowing through a pipe guides our actions, not just our language (Gentner and Gentner, 1983; Stocklmayer and Treagust, 1996; Harrison and Treagust, 2006). Metaphor is thus not a passive expression of objective similarity, but may create understanding.

While this is certainly true for a large class of metaphors and adds a genuinely new idea to Heidegger's thought, something is also wrong with this unidirectional account.

First, it does not explain the instability of similarity. Nothing in language seems to link blue eyes and green eyes anymore than it would link blue shoes and green shoes. If the similarity is not based on objective features either, then we need to explain where it comes from.

Second, it does not explain the apparent difference between live and dead metaphors. For instance, the act of closing a container and the act of killing a creature can both function as source domain for "causing something to end," but there is a notable difference in freshness within each of the following pairs:

(4.14) (a) I closed my company.
(b) I killed my company.

(4.15) (a) I closed ignition process.
(b) I killed the ignition process.

If metaphors were really responsible for creating their own similarity, we should expect them to create equally much.

From the Heideggerian point of view, a metaphor is an expression of something "as" something. Principally, it is therefore an expression of a non-verbal understanding of the life-world. On the other hand, our existence is open to new ways of relating to the world, and our fellow human beings can certainly cause such shifts. Specifically, hearing someone talk in an alien way may cause us to relate to the world differently. Since things appear differently to different existential roles, this will change our perception of similarities between things.

For instance, the reason that we can *spend out time* or *use two hours of labour* is that our culture measures and plans time. This is a social fact that we have a extensive first-hand experience of. Actually, an important function of school is to teach us to comply with the structures set up by timetables and clocks. The schedule on the bulletin board and the sound of the school bell are props that makes this abstract social fact tangible. They are, in Heideggerian terms, equipment.

A metaphor like TIME IS A WHEEL, however, is not a part of our cultural inventory. If we were exposed to stories and pictures that invited us to see the world in terms of this metaphor, we might either do so or reject the alien perspective as nonsense. Either way, there would be an intermediate period in which our own understanding was perturbed, causing a feeling of insecurity or, in existentialist terms, "anxiety" (cf. Kierkegaard, 1980; Perkins, 1985).

In smaller measure, this occurs whenever we are exposed to an action–context pair not entirely familiar to the language-game we are playing. Sentences like *I killed my company* or *I closed the ignition process* are neither completely unintelligible nor completely standard. They consequently have an air of novelty that we may be attracted to or repelled by. In other words, this analysis explains the freshness of a novel metaphor as the freshness of a novel understanding.

4.3 Bidirectional metaphor

The reinterpretation of the phenomenon of metaphor that I proposed in the previous section ought to make us ask which conceptual reservoir we draw metaphorical meaning from. In *Metaphors We Live By*, Lakoff and Johnson propose that the foundation under our concepts is "simple spatial concepts, such as UP" (Lakoff and Johnson, 1980, p. 56). Their basis for this claim is the variety of examples they have discussed, many of which conceptualise abstract things in terms of movement, orientation, distance, and posture.

In their analysis, we should thus see the sentence *he took the job* as a metaphorical application of the verb *take* in its literal base meaning, that is, grasping with your hand. To underscore this conclusion, they consider the following three sentences:

(4.16) (a) Harry is in the kitchen.

(b) Harry is in the Elks.

(c) Harry is in love.

According to their theory, only the first is literal:

> The concept IN of the first sentence emerges directly from spatial experience in a clearly delineated fashion. It is not an instance of a metaphorical concept. The other two concepts, however, are instances of metaphorical concepts. The second is an instance of the SOCIAL GROUPS ARE CONTAINERS metaphor, in terms of which the concept of a social group is structured. (Lakoff and Johnson, 1980, pp. 59–60)

This asymmetry is a metaphysical assumption, though. Empirically speaking, they might as well have said that *in the kitchen* was a metaphorical extension of *in love*. There are plenty of metaphors in which the vehicle is not a spatial concept, nor even a concrete experience: ARGUMENT IS WAR, TIME IS MONEY, IDEAS ARE COMMODITIES, or LOVE IS MAGIC (cf. Lakoff and Johnson, 1980, ch. 10).

This is not a problem unless we cling to the assertion that all these concepts (WAR, MONEY, COMMODITIES, MAGIC, etc.) should derive their meaning from bodily or spatial experiences. If we do claim that, we are forced to assume that there is a hidden conceptual structure in our concept of WAR, grounding it on knowledge of physical force, movement, orientation, and the like. However, that is nearly as implausible as assuming that everyone knows a definition of every word they have ever used, and much for the same reasons.

There is a more economical, albeit less theoretically grandiose, way to explain these linguistic forms. Lakoff and Johnson hint at it themselves:

> Even if you have never fought a fistfight in your life, much less a war, but have been arguing from the time you began to talk, you still conceive of arguments, and execute them, according to the ARGUMENT IS WAR metaphor because the metaphor is built into the conceptual system of the culture in which you live. (Lakoff and Johnson, 1980, pp. 63–64)

The concept of WAR does thus not have to be grounded directly in experience of war. This points in the direction of a much larger base level of meaning: Anything which has a designated cultural role is bestowed with a meaning. We can, as it were, know about WAR without necessarily knowing about war. Consequently, many more concepts are left unanalysed, for better or for worse. Jobs, insurance, money, gambling, driving, medicine, holidays, housing, and politicians are all a part of the everyday world, even though they are not bodily or spatial concepts.

But the simplest theory is not always the best. The Heideggerian theory of metaphor removes the problems with, for instance, bidirectional conceptualisations such as these (taken from various internet sources; in some cases slightly simplified):

(4.17) (a) The founding *fathers* found voting important. (POLITICAL LEADERS ARE PARENTS)

(b) My mom's a *dictator*. (PARENTS ARE POLITICAL LEADERS)

(4.18) (a) There's an ugly *edge* to that joke. (A SPEECH ACT IS AN OBJECT)

(b) This knife is a *joke*. Don't buy it! (AN OBJECT IS A SPEECH ACT)

(4.19) (a) It's like my only *lover* is my dope. (A DRUG IS A LOVER)

(b) She's my *heroine*. (A LOVER IS A DRUG)

(4.20) (a) This car should *convince* you to love the 1973 GTO's! (A CAR IS AN ARGUMENT)

(b) Barry's arguments *ran into a wall* of skepticism. (AN ARGUMENT IS A CAR)

So *lover* may be conceptualised in terms of *drug*, for instance, but *drug* may also be conceptualised in terms of *lover*. If our experience of the everyday world was really ordered in a strict hierarchy, at least one sentences in each pair would be unintelligible or unnatural. However, drugs, parents, cars, lovers, jokes, and arguments all lie well within the sphere of average everydayness, and they all have a designated role in our everyday social encounters.

And the list could go on. Concepts like *foreign*, *employ*, *inherit*, *professor*, *kindergarten*, *guarantee*, *ticket*, *road map*, *Hitler*, or *money in the bank* do not designate concrete bodily or spatial experience, but we all have a perfectly good sense of what they are. The claim that all of these concepts should derive their meaning from some underlying conceptual structure rooted in spatial and bodily experience is based on an unwarranted extrapolation from metaphors like *ahead of his time* and *a hot topic*.

The alternative is to see metaphorical meaning as a derivate of our intuitive understanding of the world. The instability of similarity and the differences between live and dead metaphors can then be explained in terms of the freshness of the understanding rather than the freshness of the expression.

As a consequence of this expansion of the base level, we would have to formalise our entire everyday experience in order to cope with metaphorical meaning production, not just our spatial and bodily experience. Since life is subject to existential openness, so is our conceptualisations. Metaphorical meaning production will thus continue to create disambiguation problems for machine translation, not just one or two decades ahead, but indefinitely.

4.4 Metonymy

Metonymy is traditionally defined as the substitution of one phrase with another whose "meanings are near each other in space or time," as Leonard Bloomfield puts it (Bloomfield, 1935, p. 427). The classic examples are sentences like *The Hague demands help from Interpol* and *He works at the newspaper*, where *The Hague* and *the newspaper* do not refer to the city and the publication, but to institutions associated with them.

Such substitutions are abundant in everyday language. Some examples are, in roughly increasing order of lexicality,

(4.21) We will run out of gas before *Reykjavík*. (= the time we reach reach Reykjavík)

(4.22) Jessie said she heard *the explosion*. (= the sound of the explosion)

(4.23) He has had a number of flashbacks since *Vietnam*. (= the war in Vietnam)

(4.24) He died in *9/11*. (= the terrorist attacks on 9/11)

(4.25) He stood in *the door*. (= the doorway = the opening usually blocked by the door)

(4.26) We cut the *legs* off of our pants. (= the parts that cover the legs)

(4.27) I have *the cold*. (= the disease you get from exposure to cold)

(4.28) I need a *teaspoon*. (= a spoon typically used when drinking tea)

Note that in all of these cases, the literal reading of the sentence is at best wrong and mostly nonsense: You can, for instance, not do anything *since* a country. We are consequently forced to read the *Vietnam* in *since Vietnam* as a "reference point" merely pointing towards an actual "target," in this case the war in Vietnam (Langacker, 1999, ch. 4.3).

But there is more to metonymy than closeness in time or space. Many other types of connections can form the basis of a metonymy, for instance cause–effect, producer–product, controller–controlled, signifier–signified, type–instance, and part–whole. The following examples illustrate this:

(4.29) The floor had *cigarette burns* all over. (= marks caused by burning cigarettes)

(4.30) Those are *cherry blossoms*. (= the blossoms of *a cheery tree* = a tree that bears cherries)

(4.31) This is still within *the letter of the law*. (= a strictly literal interpretation of the law)

(4.32) I wrote *that book*. (= the work that that particular object embodies)

(4.33) *Winston Churchill* won the war. (= the forces under the command of Winston Churchill)

(4.34) The company may terminate *its plans*. (= the series of actions described in the plans)

(4.35) That sounds like *a car trying to start*. (= someone trying to start his or her car)

(4.36) Gordon Brown has appointed some new *faces* in Government. (= people with faces)

None of the relationships mentioned above necessarily require closeness in time or space. The manager of a company can easily be both a cause, producer, and controller without ever being close to his effects, products, or controlled subjects.

One of problems with metonymy from the point of view of language processing is that not all kinds of relations seem equally relevant from the point of everyday human life. An employer can hire *hands* for an assembly line because hands and people appear close in the context of the assembly line. He cannot hire **feet* or **livers*. Plenty of present-at-hand relations are consequently disqualified as reference points in many contexts:

(4.37) He closed *the doorway.

(4.38) I have *the bacteria.

(4.39) Jessie said she heard *the flash of light.

(4.40) Have you seen today's *journalists?

(4.41) Gordon Brown has appointed *some new eyeballs in Government.

Ronald Langacker interprets the mechanism behind metonymy as a "shift in profile" (Langacker, 2008, p. 250): If the thought of one thing (like My Lai) automatically evokes the thought of another (like the My Lai massacre), the former can be used as a reference point for the other (as in *since My Lai*). Langacker illustrates this graphically with a diagram similar to the one shown in figure 4.1.

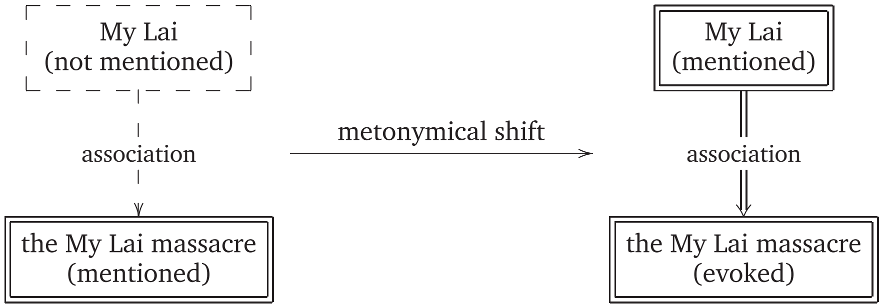


Figure 4.1: An analysis of the metonymical shift from *since the My Lai massacre* to *since My Lai*, based loosely on Langacker (2008, p. 250).

Since not all associations are experienced as relevant, the links may be symmetrical or asymmetrical. You can *run out of gas before Reykjavík*, but you cannot **be heading for in two hours*, even if you will reach Reykjavík in two hours. Similarly, you can *behave weirdly since Iraq*, but not **go to the Iraq war*. That one thing evokes another does not guarantee the that the reverse is true.

Similarly, not all associations as perceived as equally strong. The terrorist attacks on September 11th, 2001 can be referred to by their date, but the assassination of John F. Kennedy cannot. This is a matter of contigent real-life parameters like the fact that the attack on September 11th occurred in several places at once, and that no one took the blame for it. There was thus no obvious event–place or cause–effect relationship was available for reference, even though these are usually preferred over event–time.

Highly lexicalised metonymies like *teaspoon* or *cherry tree* might create the impression that metonymy is a simple matter of picking the "obvious word" when referring to everyday things and events. But since practical life is not a matter of rule-application, there are often several conflicting ways to produce and interpret a metonymy.

For instance, the standard grammatical pattern in English is not the one between tea and a teaspoon. Normally, a modifier noun refers to something a modified noun contains or consists of, like in *olive oil*, *bronze statue*, *Arab state*, *ice cube*, *anarchist movement*, *silk dress*, *brick house*, etc.

But other types of relations are equally productive. The instrument–purpose relation in *teaspoon* is also found in *baby oil*, *emergency plan*, *car pool lane*, *television set*, *freight train*, *beer crate*, *milk carton*, *school teacher*, *finance minister*, etc. This is an actual grammatical difference: If you take the bronze out of a bronze statue, it is not a *bronze statue* anymore, but if you take the freight out of a freight train, it may still be a *freight train*.

This gives rise to openness and ambiguity. A *clay oven* may be either an oven made of clay or intended for clay. In some cases, yet another relationship will have to be invoked to explain the construction: A producer–product relationship in the case *news network*, *car factory*, *computer output*, *court rulings*, and *screenplay writer*, or a part–whole relationships in the case of *car radio*, *train cart*, *farmhouse*, and *church tower*.

It is extremely hard to say anything general about the boundaries between these classes for much the same reasons that it is hard to enumerate all language-games. There is nothing in language itself that can tell you why *baby oil* and *olive oil* fall in two different categories—that is just a matter of knowing what normal people normally do, and how they normally do it. It is not written down anywhere that a *teaspoon* does not produce or consist of tea, but because of our familiarity with everyday life, we do not even consider the option. We just intuitively respond the appropriate way at the appropriate time, as if a ball thrown at us.

In other words, the apparent unambiguity of these words is a product of the overlap in the way a speech community relates to their everyday world. In order to understand how such a community uses metonymy, we have to have a deep understanding of what kinds of relevances and connections they see in everyday phenomena. When we see the natural connection between countries and wars, or between fevers and beds, we are essentially exhibiting a piece of anthropological knowledge.

This becomes particularly evident in cases where the connection between two phenomena are very strong, like an explosion and the sound of an explosion. Consider, for instance, the following sentences:

(4.42) (a) I heard someone playing the violin.
(b) I heard someone's violin.

(4.43) (a) Most of the electorate was unhappy with this procedure.
(b) Most of people in the electorate were unhappy with this procedure.

(4.44) (a) The car turned around.
(b) The driver turned her car around.

These sentences can obviously be read so that they describe the same situation. On that reading, at least one of them will strictly speaking have to be metonymical: If we agree that a car cannot literally *turn around*, then (4.44 a) must be derived metonymically from (4.44 b), but if we agree that it is actually the car that turns with the driver only causing it to turn, then we have the opposite relationship. By the same token, we can either decide that a violin have a sound of its own, or that it is actually the violinist coupled with the violin that makes the sound. In an abstract meaning representation, these options would show up as different as selling a gun and committing a murder.

The point is that they do not appear that way in everyday life. For all practical purposes, we would be making the same moves in the same language-games if we decided to stop saying *I hear a violin* unless we heard a violin playing without help from any person. We might be able to conceive of some theatrical contexts where *hearing the explosion* and *hearing the sound of the explosion* are two different things, but it takes a lot of violence to the constraints of everyday normality.

It is thus no coincidence that our language games draw their maps the way the do. If the language-users do not care about a certain distinction, no one else will do it for them. Language is not a tool for objective description, but for social interaction between people with a largely shared phenomenal world. Once we start making distinctions between anything that might ever be relevant to any bizarre creature, there is no end to the process.

Let me just conclude with a list of English-to-German mistranslations, that may be seen as due to misinterpreted metonymies. They were produced by Google Translate on December 21st, 2009:

(4.45) Let's go to *your place*. (= the place where you live)
Lassen Sie sich *Ihren Platz* zu gehen. (= your spot, location)

(4.46) It doesn't look like his *writing*. (= handwriting)
Es sieht nicht so aus wie sein *Schreiben*. (= act of writing)

(4.47) Were you at *her birthday*? (= her birthday party)
Warst du *bei ihr Geburtstag*? (= "with, adjacent to" her birthday)

(4.48) He died in *9/11*. (= the terrorist attacks)
Er starb im *Jahre 9 / 11*. (= "the year 9 / 11")

(4.49) I have *the cold*. (= the disease caused by low temperatures)
Ich habe *die Kälte*. (= the low temperature)

(4.50) The diplomat talked *to Washington*. (= the people in Washington)
Der Diplomat sprach *in Washington*. (= in Washington)

Again, some of these examples also feature errors not due to metonymy misinterpretation, but I have ignored that as far as possible.

4.5 Sound iconicity

Most laypeople are aware that there are onomatopoeic or "sound-imitative" words in their language, like *cuckoo* or *buzz*. However, standard examples like these seem to suggest that the phenomenon is limited in scope, and that the sound icons of a language might easily be singled out and enumerated.

That is not an accurate picture. As the examples in table 4.1 show, the practice of mimicking sounds in writing is productive and can give rise to indefinitely many new, but intelligible words. *why is [User 37] grrrrrring* or *I am zzzzzzzz* were not linguistic constructions before the anonymous users from the NPS Chat Corpus invoked them, but we have no problems coining and understanding such neologisms. A human translator with a normal experience of how things and people sound would also have little trouble translating them—even if they would have to be changed in form, for instance into the Dutch *waarom grrrrrrt hij?* or the Danish *jeg zzzzzzzzer*.

Machine translation programmes, on the other hand, give up confronted with most of these non-standard expressions. For instance, English-to-German translations of the sentences in table 4.1 quicly show a discouraging pattern (in translations done by Google Translate on December 21st, 2009):

(4.51) (a) awww [User 35] yer gonna be what? 57 richtig?
(b) i hate you [User 121] fuck your ugly
(c) sorry didnt see ya ...
(d) Vielleicht b / c I am zzzzzzzz
(e) scheint in den letzten Nächten I in hier bin ich, wie zzzzzzzing & das LOL sucks
(f) luvs Kürbis-Kuchen mit Sahne whuuped
(g) Ich kann seeee youuuu.

Another important point is that overt sound iconicity is quite common, at least in the type of prose recorded in the NPS Chat Corpus. There are literally hundreds of posts in the corpus that play on the sound of the letters and the sound of real-world phenomena. Without a proper understanding

October 19, adults, entries 382 and 392:

```
User 35: Ugh, my b-day is in 5 days =(
User 32: awww [User 35]  yer gonna be what?    57 right?
```

October 19th, 20s, entries 315, 331, 335, and 336:

```
User 7: i hate you [User 121] fuck your ugly
User 121: I am ugly?
User 121: aww.
User 121: i feel bad now.
```

October 24, 40s, entries 594, 604, 613:

```
User 37: grrrrrrrrrrrrrrrrr
User 50: why is [User 37] grrrrrring?
User 37: damn puter
```

November 8, 40s, entries 490, 493, and 653:

```
User 48: sorry didnt see ya...
User 48: maybe b/c I am zzzzzzzz
User 48: [User 66]............ seems the last few nights I
         been in here I am like zzzzzzzing & that sucks LOL
```

November 9th, 20s, entries 1, 5, and 7:

```
User 59: im like five seconds from ripping into a bottle
         of jack
User 59: tick tock
User 59: tick tock
```

November 9th, 40s, entries 231, 232, and 236:

```
User 18: luvs pumpkin pie with whuuped cream
User 19: mmmmmm [User 18]
User 8: mmmmmmmmmm [User 18] with whipped creme?
```

November 11th, teens, entry 349:

```
User 111: I can seeee youuuu.
```

Table 4.1: Examples of sound iconicity from the NPS Chat Corpus (Forsyth and Martell, 2007). The corpus was compiled in 2006 by Eric Forsyth, Jane Lin, and Craig Martell, all from the Naval Postgraduate School in Monterey, California (cf. `http://faculty.nps.edu/cmartell/NPSChat.htm`). It is divided into parts named after they day they were recorded and the target age of the chat room. The user names are masked for anonymity.

of both, an entire channel of linguistic information will be blocked out. Let me try to show that in a little more detail:

In an experiment done in the early '90s, the ehtnobiologist Brent Berlin compiled a list of words for "squirrel" and "tapir" in 19 indigenous languages spoken in South America. He arranged these words in pairs of which one was a word for "squirrel," and the other one was the word for "tapir." When he asked a class of students to guess which was which, they had an overall success rate of 79% (Berlin, 2003).

This is impressive, but not at all surprising when you look at the list of words. *mezaha/kuzikuzi*, *baaetae/titi*, and *tsotinba/tiriri* are not random aggregation of phonemes. /i/ is a relatively light vowel compared to /a/ or /o/, and it is often overrepresented in words that concern a certain fuzzy set of qualities, including high speed, small size, and the like. The set of synonyms for *tiny* in WordNet 3.0, for instance, consists of the words *bantam*, *diminutive*, *flyspeck*, *lilliputian*, *midget*, *petite*, and *tiny*—5 out of 7 of which have an /i/ syllable.

This association of a vowel sound with a set of characteristics seems so natural that it is hard to appreciate—a small bell just says *pling*, and a large one says *dong*. But that is only because we draw on a rich experience of how it feels to pronounce a phoneme and of how things and people sound. In order to catch this non-linguistic first-hand experience, we would have to keep a record of every sound in the world that might be relevant, represented in a format that preserves any distinction that might be relevant.

The tapir/squirrel experiment is telling, but it is also just a probe with a small sample size. In a more recent experiment, Berlin gave 600 subjects a list of 50 words in Huambisa, a small language spoken by an indigenous people of Peru (Berlin, 2006). The words were ordered in pairs like *chunchuíkit/máuts*, where one of the words was a bird name, while the other was a fish name. The subjects where then asked to check the word they thought referred to a bird.

The 100 sets of answers that Berlin had examined when he wrote his report showed that the students were able to identify the unfamiliar bird names with a 58% accuracy level.[1] He remarks:

> This level of accuracy is not high, but is much better than chance and is extremely significant statistically ($p = 0.005$) given the large number of comparisons involved, in this case exactly 5000. (Berlin, 2006, p. 78)

For some pairs, the accuracy was even higher: 98% of the subjects were able to tell that *chunchuít* is the bird and *máuts* is the fish. Remember that this is in a language that the student had no prior knowledge of—they did not even know any languages from the same family.

I hope this makes it clear that sound iconicity is a very real phenomenon that gives rise to an entire dimension of meaning in addition to signification by convention. Machines are not embedded in the same meaningful sound universe we human beings are, and it seems unlikely that the mechanism behind sound iconicity can be formalised. This mechanism thus constitutes yet another channel of information that we can employ when translating a text, but computers cannot. In the next subsection, I will try to gauge the depth of this problem in a little more detail.

Since the concept of sound iconicity is fairly representative, I will not discuss its two cousins, time and space iconicity, here. It is worth noting, however, that linebreaks, dashes, spaces, parentheses, capital letters, ellipsis marks, punctuation marks, indentations, etc. can play on familiarity

[1] Of course, this does not mean that the students would have been able to guess the exact meaning of the words without proper context—but it does mean that the proper context in itself does not provide all the information. The meaning of an utterance such as *You are such a slaz!* can roughly attributed to two sources, the frame *You are such a _____!* and the sound and shape of my newly invented word *slaz*. Had I kept the frame but instead filled it by the word *gronk*, *doop*, or *teep*, the meaning would be different.

with the spatial and temporal world in ways that are presumably as hard to formalise as the play on sound familiarity is. For further justification and examples, the book series *Iconicity in language and literature* is a valuable source (Nänny and Fischer, 1999, 2001; Müller and Fischer, 2003; Maeder et al., 2005; Ljungberg and Tabakowska, 2009).

4.6 Derived sound iconicity

If you think of sound iconic words in terms of mimicry, you would expect most of them to refer to sounds. This is certainly so for a number of them, like *hiss, buzz, sizzle, shoo, swoosh, squeak, knock, bang, crash, puff* and *splash*. But it is remarkable that practically none of these words refer exclusively to a sound: You can, for instance *give someone a buzz* as well as *hear a buzz*, but not **give someone and hear a buzz*.

This is the trace of a metonymic mechanism at work. If you need to refer to the act of "making liquid strike or fall on something in drops," the sound of water is a handy reference point to invoke. *splash* consequently has one meaning denoting a sound, and another one denoting something intimately related to the sound. (Both meanings have metaphorical extensions, too, but that is not what I want to discuss here).

What about less intimate relations? There may be a more general iconic mechanism in language, which attaches meaning to words on the basis of even weaker sound relations. Would words like *grip, grasp*, and *grab* work the way they do if they sounded, say, *doode, fliff*, or *sorve*?

And what about the similarity of a word to another? For example, according to the *Oxford English Dictionary*, the English word *obnoxious* was until the 19th century used in the neutral sense of "liable, subject, exposed, or open to a thing (esp. something actually or possibly harmful)." However, its similarity to the pejoratives *obstinate, obstreperous, obscene, obscure, oblivious, obdurate*, etc. seems to have caused it to acquire the pejorative meaning "offensive, objectionable, odious, highly disagreeable" (Fischer and Nänny, 1999, p. xxiii and Simpson, 2009, *obnoxious*, senses 1 and 5). Similar observations hold for other pairs, like *flout* and *flaunt*.

Once we start thinking in terms of these mechanisms, a lot of words suggest themselves as possible results of subtle sound iconicity. Is it a coincidence that the words *ice, snow, freeze, frost, shiver*, and *frigid* all employ /s/- and /f/-sounds, or does the sound of a freezing person play a role? Would the word *goofy* sound equally silly if the words *goon, loony*, and *fool* sounded different? And what about *trip, topple, stumble, punch, fluffy, bit, tiny, little, nanny, mommy, chatter, twitter, prick, point, snooze, doze, daze, drowse, tip, tap*, and *top*?

These are good, empirical questions that should not be answered with a few casually strewn examples. A lot of individual cases and types of iconicity have been examined in *Iconicity in language and literature*, but for the rest of this subsection, I will focus on one particular example, namely the English words beginning with *gl-*, like *gland, glossy, glove*, and *glue*. Piotr Sadowski (2001) has investigated this group of words and found strong evidence of meaning transference mediated by the common *gl*-sound.

Sadowski's study uses the spelling of the words as entry point rather than their actual pronunciation. However, we should remember that the similarity in sound might not always be as easily read off the spelling, as the above-mentioned *ice/freeze/shiver* examples indicate. In the case of the *gl*-words, though, the orthography works fairly straightforward, and I will assume that the results found for this word group generalise to more orthographically obscured sound similarities.

The method Sadowski used included an inspection of the entire set of *gl*-words listed in the

Oxford English Dictionary. He divided this large set into 46 classes with distinct etymological origins. This conflated words like *glaze* and *glass* because they are historically derived from the same source. After the etymological clustering, a representative was chosen for each of the 46 classes. The result was 46 representatives: *glad*, *glare*, *gloom*, *glory*, *gloss*, *gluttony*, etc.

Then the dictionary entry for each representative was examined, and the set of definitions in the dictionary entries were reduced to a coarser set of meaning labels. The representatives *glimmer* and *glitter*, for instance, both have meanings concerned with light, so they were put in the class labelled "light, brightness." The 10 meaning classes and their members are given in table 4.2.

	Meaning	Words	Number
1	Light, brightness	*glad* (8), *glade*, *glaik* (4), *glance* (2, 3), *glare* (2), *glass*, *gleam*, *glee* (8), *gleed*, *gleg* (6), *glent* (2, 3), *glimmer*, *glisten*, *glitter*, *glow*	15
2	Looking, seeing	*glance* (1, 3), *glare* (1), *glent* (1, 3), *glint* (3), *gloat*, *gloom* (5), *glower* (5), *glut*	8
3	Moving lightly	*glace*, *glaive*, *glance* (1, 2), *glent* (1, 2), *glide*, *glint* (2), *glissade*	7
4	Deceiving	*glaik* (1), *glaver*, *gleek*, *glib* (6), *gloze*	5
5	Dark light	*gloaming*, *gloom* (2), *glower* (2), *glum*	4
6	Smoothness	*glaborous*, *gleg* (1), *glib* (4), *glossy*	4
7	Slimy substance	*glair*, *gleet*, *glue*	3
8	Joy	*glad* (1), *glee*, (1)	2
9	Splendour	*glamour*, *glory*	2
10	Miscellaneous	*glack*, *glen*, *gladiator*, *gland*, *glean*, *glebe*, *gleg* (prudence), *gloss*, *glove*, *gluttony*	10

Table 4.2: A semantic categorisation of *gl*-words in English, taken from Sadowski, 2001, p. 76. I have added parentheses with references to other semantic classes when a word falls into more than one of the named categories.

As I will argue in more detail in a minute, this inspection shows an extraordinary degree of relatedness between the English *gl*-words. Almost all of the etymologically distinct words can be grouped into a few meaning categories, and polysemy across the category boundaries is frequent. In addition to the semantic proximity between the words, the meanings are also thematically connected: The real-world phenomena of "looking" and "light" are in a phenomenal sense closer than, say, "looking" and "food". Since the words were picked purely on the basis of their two first letters, this gives us very strong evidence that (some types of) similarity in sound is correlated with (some types of) similarity in meaning.

Sadowski himself certainly agrees that the meanings of the *gl*-words are packed more densely than we should expect from a random sample and supports it by pointing to the relatively small size of the "miscellaneous" group: Only 10 of the 46 word-classes on his list "do not appear to be related in a direct way to any of the distinguished semantics groups" (Sadowski, 2001, p. 76). That is true, but we should remember that the grouping is based on a set of *ad hoc* meaning labels that Sadowski created for the occasion. His argument would carry more weight if he had measured the compactness of the groups of word-classes by some more general method.

In order to do this, I devised a rudimentary measure of compactness based on the dictionary entries in WordNet 3.0 (Fellbaum, 1998). With the help of a bit of statistics, this measure showed that the *gl*-words are packed with a density far above chance level.

The concept of density only makes sense if we have a metric on the meaning space, so I needed a measure of the distance between two word meanings. There are obviously many ways to measure semantic distance (Leacock and Chodorow, 1998; Lesk, 1986; Lin, 1998; Resnik; Wu and Palmer, 1994), but I wanted to define a simple method that would catch the elusive tie between, say, *screw* and *screwdriver*. Since such words are not really close in terms of the logical relations recorded in WordNet (like membership and entailments relations), standardised measures based on such relations also fail to catch the connection. The Wu-Palmer measure (Wu and Palmer, 1994; Pedersen et al., 2004), for instance, rates the similarity of "`screw (noun) 4`" and "`screwdriver (noun) 1`" as 0.600, but gives "`screw (noun) 4`" and "`remote_control (noun)`" a higher score, 0.778. This seems particularly ridiculous in view of the WordNet definition of "`screwdriver (noun) 1`": "a hand tool for driving screws; has a tip that fits into the head of a screw."

In other words, a measure not based on logical relations was called for. I therefore ran through WordNet's dictionaries, split the definition of each entry into its component words, and added the word's synonyms to that list. This gave me a list of "features" for every meaning of every word in the WordNet database. The feature lists are of course idiosyncratic to the particular definitions in WordNet, but it is hard to see how that idiosyncrasy would in any way be particularly biased towards relating the *gl*-words if not because of similarities in meaning.

When measuring the similarity of two meanings, I count how many features they have in common, counting features multiple times if they occur multiple times on one of the lists. By dividing this integer by the number of possible matches (product of the two list lengths) I get a ratio I can use as similarity score. This procedure is illustrated graphically in figure 4.2.

This similarity score is very sensitive to the length of the definitions, and it does not account for the fact that a feature like `"the"` is much more common than a feature like `"brittle"`. This has some bizarre consequences that would disqualify the similarity score as a all-purpose semantic proximity measure. However, this bias is irrelevant for my purposes, so I have prioritised the statistical simplicity of the measure over its general applicability.

Unfortunately, some of the *gl*-words on Sadowski's list do not have dictionary entries in WordNet, so in my calculations, I had to leave out the words *glaborous*, *glack*, *glaik*, *glair*, *glaive*, *glaver*, *gleed*, *gleek*, *gleg*, *glent*, and *gloze*. That leaves 35 words. Most of these words have several meanings recorded in the database, and I compiled a set of 199 word senses with different numbers of features.

There are $199 \times 199 = 14{,}161$ pairs of such feature lists. I found their average pairwise similar-

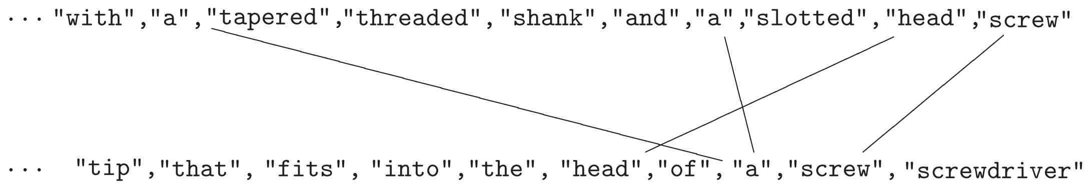


Figure 4.2: The word senses "`screw (noun) 4`" and "`screwdriver (noun) 1`" have 12 and 18 features, respectively. In a pairwise examination of the features, we find 11 identical cases of identity among the 216 pairs. The similarity score for the two word senses is thus 11/216 or approximately 5.093%, which is relatively high.

Word sense 1	Word sense 2	Score
`glare (noun) 1`	`glower (noun) 0`	20.000%
`glee (noun) 1`	`gloat (noun) 0`	20.000%
`glare (verb) 0`	`glower (verb) 0`	12.500%
`glance (verb) 0`	`glint (verb) 1`	12.500%
`gleam (noun) 1`	`glimmer (noun) 0`	12.000%
`gleam (verb) 1`	`glimmer (verb) 0`	12.000%
`gleam (noun) 0`	`glow (noun) 6`	11.111%

Table 4.3: Some examples of high similarity scores between senses of *gl*-words.

ity score to be 0.948%, in that the two compared features were identical in 15,503 of the 1,635,841 cases.

The similarity scores range from 0.0% to 33.3%, or from 0.0% to 23.3% if we exclude cases in which one meaning of one lemma is compared to itself. Distinct meanings of a single lemma also tend to achieve a high similarity score. The different senses of *glass*, for instance, are among the most similar of the 14,161 pairs. Other than these obvious examples, similarities were also found between words like *glare* and *glower*. I have given some examples of the similarity scores in table 4.3.

The question is whether these scores are generally higher than we might expect from a random sample of words. In order to show that, it is necessary to do a little statistics.

If we measure the density of a group of words, the feature list of every word must be compared to the feature list of every other word. When we compare two feature lists, every feature on one lists must be compared to every feature on the other.

Thus, once we have created and paired all the feature lists for a group of words, we have established a long list of feature pairs that are to be compared. At that point, we can in practice forget about the word senses and only look at the pairs of features that we are going to compare. If we think of these feature pairs as a sample from a larger population, it is natural to model their behaviour as a binomial distribution, in this case with an estimated parameter of 0.948%.

Word sense 1	Word sense 2	Score
build (verb) 7	microbic (adj) 0	6.818%
build (verb) 3	augmentation (noun) 2	6.667%
microbic (adj)	0 equivocal (adj) 0	6.220%
gin (verb) 1	chisel (noun) 0	6.153%
build (verb) 8	give_up (verb) 11	6.000%
precede (verb) 2	augmentation (noun) 2	5.714%
underneath (adv) 0	augmentation (noun) 2	5.454%

Table 4.4: Some examples of high similarity scores between word senses in the control group.

The 5% confidence interval around this parameter can then be approximated using Wilson's method, which gives the interval [0.932%, 0.962%]. In other words, if we drew a new sample of feature pairs from the same distribution, we would with 95% certainty see the percentage of matching features lie within this interval.

These numbers should be contrasted with the density of WordNet as a whole. WordNet's dictionaries are too large to allow for a direct measurement of the similarity scores of all word senses, but we can estimate the density with two different methods: One way is to estimate the overall density by computing the general probability of two features being identical; another is to randomly pick 35 words with 119 senses, and then find their density score. I have used both methods.

In order to compute the probability of randomly picking two identical features, I have assembled the list of features from the feature lists of all dictionary entries in WordNet. Using this list, I can determine the probability of randomly picking any single feature, say, the string `"head"`. By squaring this probability, I find the probability of a randomly selecting the couple (`"head"`, `"head"`). If I sum all such probabilities, one term for each possible feature, I get the general probability of picking two identical features. I find this number to be 0.705%.

The other method is quite straightforward: I take the set of all all dictionary entries in WordNet, and for each of the 35 *gl*-words, I pick a random control word with the same number of word senses. This gives me 119 word senses with 199 feature lists. I can then proceed to compute their density the same way I computed it for the *gl*-words. For the particular sample my computer picked, the score turned out to be 0.869%. I have shown the score for some of the pairs of word senses in table 4.4, again leaving out comparisons of senses of the same word.

Evidently, both methods indicate that WordNet's dictionaries are generally less densely packed than the *gl*-words. Both estimates lie outside the 5% confidence interval around the density score for the *gl*-words. Actually, they are even far below the left edge of the 0.1% confidence interval (which is around 0.923%), so the difference is extremely significant. We can hence conclude that the *gl*-words are remarkably close in meaning, despite the fact that they were selected purely on the basis of their sound.

There is thus a strong semantic tie between the English *gl*-words. For one part, this may be due to the fact that the first-hand experience of hearing or saying /gl/-sounds somehow reminds us of the first-hand experience of reflections of light and slippery surfaces. For another part, it is also very likely due to the fact that the meaning of words with similar sounds tend to rub off on each other. By both explanations, the linguistic sounds *glass*, *glide*, *glare*, and *glance* are not randomly picked, but close to each other because their meanings are.

If this is true, then the ability to recognise sounds—both linguistic and real-world sounds—is highly important to our capacity for expression and understanding. Sounds do not come without commitments, even though these commitments are sometimes so subtle and complex that we ourselves have a hard time putting our finger on them. We do not normally think of the acoustic quality of words like *swoon*, *cog*, *flail*, *pocket*, *woozy*, *tinker*, *haze*, *creepy*, *pub*, or *bucket*—but as we have seen in the case of the *gl*-words, that does not necessarily mean that their sounds play no role.

As I said in the previous subsection, this evasive information channel is highly problematic from the point of view of machine translation, since it potentially requires us to include the sound of everything in the world in our programmes, as well as record how these sounds gear into our everyday emotions and activities. Even if we had begun on this task, it is hard to see when we would ever finish it, and it may thus continue to bar machines from handling of utterances in context with the same competence that we display on a daily basis.

4.7 Agentivity

Animacy poses a quite simple, but ubiquitous problem for machine translation: For instance, the German word *sie* can either refer to a thing (*it*), a person (*she*), or a collection of either (*they*), and translation systems very frequently pick the wrong candidate. We have already seen Spanish-to-English examples in section 1.4.1, and I will not repeat them here. Let me instead turn to to some more subtle phenomena which would pose a problem even for a future machine translation capable of handling more mundane animacy distinctions.

A particular quirk of the English language is that it has two different genitive forms, the *of*-genitive and the *-'s*-genitive. Which one is appropriate is determined by how person-like the noun phrase in question is (Conradie, 2001; Jespersen, 1933, §14.84). Thus,

(4.52) (a) Betty's babysitter

(b) the colour of the soil

are much more natural phrases than

(4.53) (a) the babysitter of Betty

(b) the soil's colour

This is due to the fact that *Betty* is a clearly animate noun phrase that may play the role of agent in a sentence, while *the soil* is not. There are borderline cases though, as Jespersen notes (Jespersen, 1933, §§14.84–86). A *car*, for instance, can sometimes be perceived as an animate being (*he was killed by a car*) and sometimes as an inanimate (*she sold her car*), allowing both the form *the car's* and *of the car*.

This is a problem when translating from a language like Danish where there is only one genitive form. We are then forced to translate an unmarked form like *bilens* into one of the two marked forms *the car's* or *of the car*. In order to know in which category a given thing falls in a given context, you need to understand the folk conception of causation and volition that the grammatical distinction is based on.

The distinction is easily flattened by coarse translation algorithms, for instance statistical methods that simply tend to pick the most common genitive form regardless of context. Google Translate for instance translates the Danish *byens* ("the city's"/"of the city") into the agentive form *the city's* even when it obviously plays the role of a inanimate object, as in *De var bekymrede over byens voksende størrelse* ("They were worried about the increasing size of the city," translated on June 22, 2009). That form is generally more frequent, but that does not make it universally appropriate, just like the frequency of the word *a* does not allow us to stop using the word *an*.

The distinction between person-like and non-person-like makes a grammatical difference with regard to anaphora, too. If you refer to something by means of a metonymical substitute (say, *blue helmets* instead of *soldiers with blue helmets*), the subsequent discourse normally proceeds like you mentioned the actual referent, not the reference point:

(4.54) *The first violin* has got the flu. (= the violinist)

(4.55) (a) *She* can't practise today. (= the violinist)
(b) **It*'s not going to be here. (= the violin)

However, this changes if the reference point is a person and the referent is a thing. In the following two sentences, for instance, references to the actual subject of the discourse are incoherent, and references to the substitute are coherent:

(4.56) *My ex-husband* is parked on the upper deck. (= his car)

(4.57) (a) **It* has a California license plate. (= his car)
(b) *He* is taking the bus today. (= my ex-husband)

This can easily cause problems in translation from languages like Spanish, French, German, or Dutch, where the difference between *he/she* and *it* is not marked.

The authors of the examples above, Claus Uwe-Panther and Günther Radden, suggest that the difference is due to "a general cognitive principle according to which humans take precedence over non-humans" (Uwe-Panther and Radden, 1999, p. 10). But they are well aware that there are borderline cases like the following (p. 11):

(4.58) *The harpsichord* has the flu. (= the harpsichordist)

(4.59) (a) *His* part has been taken over by the grand piano. (= the harpsichordist's)
(b) *Its* part has been taken over by the grand piano. (= the harpsichord's)

Even though (4.59 b) is entirely parallel to (4.55 b), only the former is coherent. This is due to facts about the language-game such sentences usually occur in. When a score says what *the violin is supposed to play*, or someone in a symphony orchestra refers to *the first violin's part*, we might say that they are actually talking about the person playing the violin rather than the violin itself.

But from the point of view of the orchestra, this is a quite irrelevant "grammatical" convention—in the practical dealings of an orchestra, the violin itself might as well be thought of as an agent that just needs a couple of hands to work. These two things are thus virtually interchangeable from the point of view of a certain life-form. The normal "animacy hierarchy" is consequently overridden (Silverstein, 1976; Zaenen et al., 2004).[2]

[2]Another highly interesting case occurs when the animacy hierarchy is overridden by taboo. I have found a few cases of an event being metonymically evoked by reference to a person, but only when there is reason to avoid mentioning the event itself. For instance, *I felt bad after mom* meaning *I felt bad after mom died*. Needless to say, the appropriateness or inappropriateness of such metonymies is an extremely sensitive issue and takes high social attunedness. Since these taboo metonymies are exceptional and not stock phrases occurring in the hundreds, and since different cultures have vastly different sensitivities towards such taboos, I think a statistical account would do more harm than good in this case. A set of rigid "taboo criteria" would be even worse for the reasons discussed in section 3.3.

We have now seen two grammatical distinctions affected by agentivity judgements, metonymical coherence and genitive forms. Many other examples could be given and are given by, for instance, Yamamoto (1999). Getting to grips with how person-like a thing appears is thus "of obvious importance for high quality generation and translation" (Zaenen et al., 2004, p. 119). But since it has to do with the perceived qualities of the real-world referents, it is also very hard. Let me try to show how slippery the problems actually are with some more examples.

In order to investigate the theme a little further, I have done a crude and incomplete categorisation of the genitive phrases in the Brown Corpus (Francis and Kucera, 1979). I took the 250 most common nouns from the corpus: *time*, *man*, *year*s, *way*, etc. For each word NOUN on that list, I then found all phrases of the form ARTICLE NOUN*'s* and all phrases of the form ANY-NOUN OF ARTICLE NOUN, unless the last noun was not followed by a yet another noun (such as in *town hall* or *subway system emergency phone*).

This gave me all instances of genitive phrases like *the company's* or *size of the company*. It did, however, exclude phrases like *size of the company portfolio*, *tired of the company* and *all of the company*. In order to prevent overgeneralisation, I further discarded nouns that yielded less than 10 such genitive phrases in total, leaving 62 of the 250 words.

The result of the inspection was a list of nouns, and attached to each noun, the number of times it occurred as possessor in OF-genitives and in S-genitives. By looking at the ratio between these two, I could order the nouns on a scale with the ones most likely to form an S-genitive in one end and the ones least likely to do so in the other end (cf. table 4.5). *child*, for instance, has a ratio of 6/26, so it is quite likely to form the genitive *child's*, while *night*, having the ratio 11/3, leans more towards the genitive *of the night*.

The result conformed well with the hypothesis that the more agentive nouns are more likely to form S-genitives: The S-end of the spectrum starts with agentive, animate nouns like *doctor*, *patient*, *girl*, and *child*, and the other end includes inanimate things like *stream*, *period*, *future*, and *society*. So far, an animacy-rating of new words would seem to be a good predictor of how its genitives should be formed.

But there also problems: Some of the abstract entities in the corpus form S-genitives with quite high probabilities, for instance *nation* (13/27), *earth* (7/11), or *company* (14/15). On the other hand, *society* (11/0), *universe* (16/0), and *house* (30/0) do not, even though they might reasonably be expected to fall in the same categories.

This is not accidental. In his handbook of English grammar, Otto Jespersen notes that the distinction between the two genitive forms carries meaning quite systematically for concrete nation names like *England* or *Russia*:

> When a country is thought of politically as a living being, the genitive [the S-genitive] is frequently used:
>
> > England's interest in India.
> > We should be at Russia's mercy.
>
> . . . If the country is looked upon from a purely geographical point of view, of is used: the boundaries of Switzerland, etc. (Jespersen, 1933, §14.85)

So in the concrete case of nation names, the grammatical distinction corresponds to a quite subtle difference in perception of the nation. We can observe a similar difference in meaning with examples like these:

NOUN	OF	S	NOUN	OF	S	NOUN	OF	S
doctor	1	15	*day*	21	10	*plan*	15	1
patient	3	21	*party*	13	6	*group*	20	1
girl	2	9	*water*	9	4	*house*	30	0
child	6	26	*game*	10	4	*century*	19	0
boy	4	12	*board*	11	4	*word*	16	0
town	4	11	*country*	27	9	*universe*	16	0
nation	13	27	*night*	11	3	*time*	16	0
enemy	4	8	*body*	14	3	*past*	16	0
woman	5	9	*road*	19	4	*law*	16	0
earth	7	11	*family*	19	4	*area*	15	0
man	42	45	*book*	15	3	*problem*	13	0
company	14	15	*system*	13	2	*population*	13	0
industry	7	7	*moon*	15	2	*number*	13	0
individual	11	10	*community*	19	2	*knife*	13	0
moment	6	5	*work*	11	1	*fact*	13	0
city	24	18	*subject*	11	1	*war*	12	0
pool	7	5	*room*	11	1	*stream*	12	0
state	27	17	*evening*	11	1	*period*	12	0
car	12	7	*church*	27	2	*future*	12	0
world	55	31	*river*	15	1	*society*	11	0
year	19	10	*program*	15	1			

Table 4.5: Some common Brown Corpus nouns, ordered to their tendency to form S-genitives relative to their tendency to form OF-genitives.

(4.60) (a) the occupation of the country

(b) the country's occupation

Here, we achieve two different readings on the sentence depending on whether the genitive form nudges us into looking at *the country* as an agent or a non-agent: In the first case, *the country* has a role of passive part being occupied, and in other case, *the country* takes the active, agentive role as occupier of something. The same distinction in meaning appears in many other cases, for instance *the jury's selection/the selection of the jury*, *the earth's development/the development of the earth*, or *the revolutionaries' problem/the problem of the revolutionaries*.

In order to choose the correct genitive when translating a source text without this distinction, we need come to grips with the folk theory of causation which gives *nation* (13/27), *city* (24/18), and *world* (55/31) a more agentive role than *area* (15/0), *population* (13/0), and *society* (11/0). If we do not know why or when an entity can be perceived as agentive, we will end up producing bad prose and—in cases like (4.60 a) or (4.60 b)—possibly even misleading wordings.

But the roots of this folk theory is in the structures of our everyday: The normal conception of politics is that politicians *run* our countries the way a person runs a programme or a machine, although with some common-sense reservations (for instance, they cannot turn the country off).

This conception of power and responsibility is enforced through an abundance of practices, like the use of flags in campaign material and the everyday habit of blaming politicians for all sorts of things. The perceived link between leaders and countries enables us to substitute countries for people and thus say things like *Germany was planning an attack on Poland*, letting the agentive features of the leaders rub off on the countries.

However, the word *society* does not behave that way grammatically. When we look at the way it is used and conceptualised, this turns out to be due to the fact that it is perceived more like a mathematical set than like a person. The *society*-genitives in the Brown Corpus include phrases like *the existing institutions of the society*, *members of the society*, etc. *society* thus corresponds to a different stance on fellow citizens than *nation* does—roughly speaking, *society* is correlated with the way you see them standing in a crowd, while *nation* is correlated with the way you see them enrolling in the army.

So in order to sort out how agentive *nation* and *society* may be in a particular case, you need to understand something about the way our everyday is structured and the notions of causation and volition that comes with that structure. Since there is no single source for the structure—statues of war heroes might be as relevant as newspaper articles—there is no way of capturing it without essentially recreating first-hand experience. This particular distinction thus poses yet another problem to achieving natural prose in machine translation.

Chapter 5

Conclusion

In chapter 3 of this thesis, I argued that linguistic distinctions owe their significance to distinctions in practical life. A change of eating habits can turn a count noun into a mass noun, and a power relation can turn a tautology into a threat. In order to grasp meaning, a machine translation system would thus have to grasp life.

I then argued that the dynamics of life could not be formalised or tabulated, since we human beings are open to new ways to relate to the world. When we understand what it means to *kill a company* we exhibit this openness. Since our understanding of the world is not the outcome of some kind of rational analysis, this ability to consume new meaning cannot be achieved by heaping up facts.

In chapter 4, I shifted to a more linguistic perspective on these matters. I showed how lack of hermeneutical footing makes machine translation systems fumble helplessly around with the metaphorical extensions of *touch* and the metonymical shift from writing to handwriting. These and many other phenomena are special cases of the general picture I unfolded in chapter 3.

Metaphor, metonymy, sound iconicity, etc. are thus the kind of phenomena that we should not expect machine translation systems to be able to handle, now or in the future. Since they are common and central to the meaning of ordinary prose, that deficiency will prevent fully automatic machine translation from ever attaining a human level of quality.

The errors that machine translation systems produce are not always and constantly catastrophic. Sometimes, they just arrange words strangely or pick a comical phrase. But there is no way to know whether an automatically translated sentence is a good translation of the original or whether it is one of the poisoned cookies. Machines are not in general able to know when they do things wrong for the same reasons that they are not able to do things right. Any output sentence might be the linguistic equivalent of boiling the baby instead of its milk.

Commercial and academic projects that aim at fully automatic translation should consequently expect to see little or no improvement even after another 20 or 30 years of generous research grants. For instance, documents distributed by the European Commission for external use, such as law texts or information material, are not an appropriate choice of domain for machine translation. Fully automatic translation will never achieve a quality good enough to translate such texts adequately.

On the other hand, the past 60 years of research in computational linguistics have given us some interesting results that might be applied in translation, that is, human translation. Dictionaries, corpora, translation models, language models, and parsers can be extremely useful tools for translators. If half the words of a text can be guessed correctly by a piece of software, the speed of translation can certainly be increased by the use computer applications. The case made in this thesis is just that they should know their limitations. If we spend money and energy on the development of such tools the way we have spend money and energy on machine translation until now, a significant fraction of the cost of translation might be shaved off in the future.

But as I showed in the introduction and in chapter 2, the history of machine translation has a tendency to repeat itself. The fact that we always seem to have come three quarters of the way again and again incites the hope that *this* time, we might be able to cover that last quarter. But instead of inciting hope, this situation should perhaps rather remind us of the fact that drinking a fourth beer always seems like a better idea after the first three.

And it would be appropriate to curb our enthusiasm, even though machine translation certainly is an exciting challenge. Unlike the researchers looking ahead at a promising future from half a century ago, we can no longer claim that we do not know the cost or the chances. Yet another half a century of attempts to make "machines that think, that learn, and that create" (Simon and Newell, 1958, p. 8) would indeed be resources badly spent.

Bibliography

Accipio Consulting (2006). Human language technologies for europe. Published as a part of the EU-funded TC-STAR project. Available on http://www.tc-star.org/pubblicazioni/-D17_HLT_ENG.pdf.

Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, D., Och, F., Purdy, D., Smith, N., and Yarowsky, D. (1999). Statistical machine translation: Final report. In *JHU workshop*, pages 1–42.

Anonymous (1954). Electronic translator. *TIME Magazine*, LXIII(3). Available on http://-www.time.com/time/magazine/article/0,9171,819447,00.html.

Aristotle (1902). *Poetics, translated by S. H. Butcher*. Macmillan, London. Available on http://www.gutenberg.org/etext/1974.

Bar-Hillel, Y. (1964a). A demonstration of the nonfeasibility of fully automatic high quality machine translation. In *Language and Information: Selected essays on their theory and application*, pages 174–179. The Jerusalem Academic Press Ltd., Jerusalem, Israel.

Bar-Hillel, Y. (1964b). Four lectures on algebraic linguistics and machine translation. In *Language and Information: Selected essays on their theory and application*, pages 185–218. The Jerusalem Academic Press Ltd., Jerusalem, Israel.

Bar-Hillel, Y. (1964c). The future of machine translation. In *Language and Information: Selected essays on their theory and application*, pages 180–184. The Jerusalem Academic Press Ltd., Jerusalem, Israel.

Bar-Hillel, Y. (1964d). Introduction. In *Language and Information: Selected essays on their theory and application*, pages 1–16. The Jerusalem Academic Press Ltd., Jerusalem, Israel.

Baum, L. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, 3(1):1–8.

Berlin, B. (2003). Tapir and squirrel: Further nomenclatural meanderings toward a universal sound-symbolic bestiary. In Sanga, G. and Ortalli, G., editors, *Nature Knowledge: Ethnoscience, Cognition, and Utility*, pages 119–127. Berghahn Books.

Berlin, B. (2006). Evidence for pervasive synesthetic sound symbolism in ethnozoological nomenclature. In Ohala, J. J., Hinton, L., and Nichols, J., editors, *Sound Symbolism*, pages 76–93. Cambridge University Press.

Billig, M. (1995). *Banal nationalism*. Sage Publications Ltd., London.

Bloomfield, L. (1935). *Language: Introduction to the Study of Language*. Allen & Unwin, London.

Bobrow, D. G., Condoravdi, C., Crouch, R., de Paiva, V., Karttunen, L., King, T. H., Nairn, R., Price, L., and Zaenen, A. (2007). Precision-focused Textual Inference. In *Proceedings of the Workshop on Textual Entailment and Paraphrasing*, pages 16–21, Prague. Available on http://www2.parc.com/isl/groups/nltt/papers/PARC_RTE3.pdf.

Booth, A. D., Brandwood, L., and Cleave, J. P. (1958). *Mechanical Resolution of Linguistic Problems*. Butterworths Scientific Publications, London.

Booth, A. D. and Locke, W. N. (1955). Historical introduction. In Locke, W. N. and Booth, A. D., editors, *Machine Translation of Languages: Fourteen Essays*, pages 1–14. Technology Press of the Massachusetts Institute of Technology, New York.

Brandom, R. (2008). Between saying and doing: Towards an analytic pragmatism. The 2005–2006 John Locke Lectures at Oxford University, UK. Updated version available at http://www.-pitt.edu/∼brandom/locke/index.html.

Bransford, J. D. and Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, 11(6):717–726.

Brown, P., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., and Roossin, P. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Brown, P., Della Pietra, V., Della Pietra, S., and Mercer, R. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Butler, J. (1999). *Gender trouble: Feminism and the subversion of identity*. Theatre Arts Books.

Carbonell, J. G. and Brown, R. D. (1988). Anaphora resolution: a multi-strategy approach. In *Proceedings of the 12th conference on Computational linguistics-Volume 1*, pages 96–101. Morristown, NJ, USA.

Carbonell, J. G. and Tomita, M. (1987). Knowledge-based machine translation, the CMU approach. In Nirenburg, S., editor, *Machine Translation: Theoretical and methodological issues*, pages 68–89. Cambridge University Press, Cambridge, MA.

Carnap, R. (1996). The elimination of metaphysics through logical analysis of language. In Sarkar, S., editor, *Logical empiricism at its peak: Schlick, Carnap, and Neurath*, pages 10–32. Taylor & Francis.

Charniak, E. (1972). Toward A Model Of Children's Story Comprehension. Technical Report AI-TR-266, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts. Available on http://dspace.mit.edu/bitstream/handle/1721.1/-6892/AITR-266.pdf?sequence=2.

Chomsky, N. (1957). *Syntactic Structures*. Mouton & Co, The Hague, The Netherlands.

Cody, S. (1905). *The Art of Writing & Speaking the English Language*. Funk & Wagnalls, New York.

Conradie, C. J. (2001). Structural iconicity: The English -S- and OF-genitives. In Fischer, O. and Nänny, M., editors, *The motivated sign*, volume 2 of *Iconicity in language and literature*, pages 229–247. John Benjamins Publishing Company, Amsterdam.

Dagan, I., Glickman, O., and Magnini, B. (2006). The pascal recognising textual entailment challenge. *Lecture Notes in Computer Science*, 3944:177–190.

Danmarks ErhvervsforskningsAkademi. Genvej til det globale marked – maskinoversættelse fra "bundlinje" til "bottom-line". Flyer for a January 2007 meeting on language technology. Available on `http://fuhu.dk/filer/DEA/T%E6nketanke/Sprogteknologi/2007/-Program%20%20produktpraesentation%20070110.pdf`.

de Saussure, F. (1966). *Course in General Linguistics*. McGraw-Hill, New York. Translated by Wade Baskin.

de Saussure, F. (1986). *Course in General Linguistics*. Open Court Publishing, Peru, Illinois, 3rd edition. Translated by Roy Harris.

Dempster, A., Laird, N., Rubin, D., et al. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Derrida, J. (1999). Hospitality, justice and responsibility: a dialogue with Jacques Derrida. In *Questioning ethics: Contemporary debates in philosophy*, pages 65–83. Routledge, London.

DeWalt, K. M. and DeWalt, B. R. (2002). *Participant observation: A guide for fieldworkers*. Altamira Press, Walnut Creek, CA.

Dillinger, M. and Rommel, A. (2004). Lisa best practice guide: Implementing machine translation. Technical Report 2, Localization Industry Standards Association (LISA), Domaine en Prael, Switzerland. Available on .

Dorr, B. J. (1993). *Machine translation: a view from the Lexicon*. MIT press, Cambridge, Massachusetts, USA.

Dreyfus, H. (1980). Holism and hermeneutics. *The Review of Metaphysics*, pages 3–23.

Dreyfus, H. L. (1965). Alchemy and Artificial Intelligence. Technical report, Santa Monica, California.

Dreyfus, H. L. (1992). *What computers still can't do: a critique of artificial reason*. The MIT Press.

Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Artificial Intelligence*, 171(18):1137–1160.

Dreyfus, H. L., Anthanasiou, T., and Dreyfus, S. E. (1986). Mind over machine: The Power of Human Intuition and Expertise in the Era of the Computer.

Dreyfus, H. L. and Spinosa, C. (2003). Further reflections on heidegger, technology, and the everyday. *Bulletin of Science, Technology & Society*, 23(5):339.

Evans-Pritchard, E. (1937). *Witchcraft, Oracles and Magic Among the Azande*. Clarendon Press, Oxford.

Fellbaum, C., editor (1998). *WordNet: An electronic lexical database*. MIT press, Cambridge, MA. See also http://wordnet.princeton.edu/.

Fischer, O. and Nänny, M. (1999). Introduction: Iconicity as a creative force in language use. In Nänny, M. and Fischer, O., editors, *Form miming meaning*, volume 1 of *Iconicity in language and literature*, pages xv–xxxvi. John Benjamins Publishing Company, Amsterdam.

Forsyth, E. and Martell, C. (2007). Lexical and discourse analysis of online chat dialog. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007)*, pages 19–26. IEEE.

Francis, W. and Kucera, H. (1979). *Brown Corpus Manual: Manual of information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Department of Linguistics, Brown University, Providence, Rhode Island, 3rd edition. Available on http://icame.uib.no/brown/bcm.html.

Frederking, R., Nirenburg, S., Farwell, D., Helmreich, S., Hovy, E., Knight, K., Beale, S., Domashnev, C., Attardo, D., Grannes, D., and Brown, R. (1994). Integrating Translations from Multiple Sources within the PANGLOSS Mark III Machine Translation System. In *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 73–80.

Frege, G. (1980). *The Foundations of Arithmetic: A logico-mathematical enquiry into the concept of number*. Northwestern University Press, Evanston, Illinois.

Gates, D., Haberlach, D., Kaufmann, T., Kee, M., McCardell, R., Mitamura, T., Monarch, I., Morrisson, S., Nirenburg, S., 3, E. N., Takeda, K., and Zabludowski, M. (1989). Lexicons. *Machine Translation*, 4(1):67–112.

Gentner, D. and Gentner, D. R. (1983). Flowing waters or teeming crowds: Mental models of electricity. pages 99–129.

Goodman, K. and Nirenburg, S. (1991). *The KBMT Project: A case study in knowledge-based machine translation*. Morgan Kaufmann.

Goodman, N. (1972). Seven strictures on similarity. In *Problems and Projects*, pages 437–447. The Bobbs-Merrill Company, Inc., Indianapolis.

Google (2009). *Google Translate FAQ*. Google, Inc. Retrieved from http://www.google.com/-intl/en/help/faq_translation.html on March 25, 2009.

Grossberg, S. (1988). Editorial. *Neural Networks*, 1(1):1.

Harris, Z. S. (1946). From morpheme to utterance. *Language*, 22(3):161–183.

Harrison, A. and Treagust, D. (2006). Teaching and learning with analogies: Friend or foe? In Aubusson, P. J., Harrison, A. G., and Ritchie, S., editors, *Metaphor and analogy in science education*, volume 30 of *Science & Technology Education Library*, pages 11–24. Springer.

Hays, D. G. (1967). *Introduction to Computational Linguistics*. American Elsevier, New York.

Heidegger, M. (1973). *Being and Time*. Basil Blackwell, Oxford. Translated by John Macquarrie & Edward Robinson.

Hobbs, J. R. (1976). Pronoun resolution. Research report #76-1, Department of Computer Sciences, City College, City University of New York, New York. Available on `http://www.isi.edu/~hobbs/PronounResolution.pdf`.

Hromkovič, J. (2004). *Theoretical computer science*, volume I of *Texts in Theoretical Computer Science*. Springer, Berlin.

Hülsbörner, S. (2007). Elektronische dolmetscher holen auf. *Computerwoche*. Available on `http://www.computerwoche.de/knowledge_center/software/591945/index.html`.

Husserl, E. (1969). *Ideen zu einer reinen Ph*
"*anomenologie und ph*
"*anomenologischen Philosophie*. Martinus Nijhoff, The Hague, Netherlands.

Hutchins, W. J. and Somers, H. L. (1992). *An Introduction fo Machine Translation*. Academic Press Limited, London.

IBM (1954). 701 translator. IBM press release. Available on `http://www-03.ibm.com/ibm/-history/exhibits/701/701_translator.html`.

Isabelle, P. (1987). Machine translation at the TAUM group. In King, M., editor, *Machine Translation: The State of the Art*, pages 247–277. Edinburgh University Press, Edinburgh.

Jespersen, O. (1933). *Essentials of English grammar*. Routledge.

Jones, D., Gibson, E., Shen, W., Granoien, N., Herzog, M., Reynolds, D., and Weinstein, C. (2005). Measuring human readability of machine generated text: Three case studies in speech recognition and machine translation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 1009–1012. Available on `http://-www.mt-archive.info/ICASSP-2005-Jones.pdf`.

Jones, D., Herzog, M., Ibrahim, H., Jairam, A., Shen, W., Gibson, E., and Emonts, M. (2007). Ilr-based mt comprehension test with multi-level questions. In *Proceedings of the conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT), Companion Volume*, volume 23, pages 77–80, Rochester, NY. Association for Computational Linguistics.

Jurafsky, D. and Martin, J. H. (2008). *Speech And Language Processing*. Prentice Hall.

Kaplan, R. M. and Bresnan, J. (1982). Lexical-Functional Grammar: A Formal System for Grammatical Representation. pages 173–281.

Kawai, M. (1965). Newly-acquired pre-cultural behavior of the natural troop of Japanese monkeys on Koshima Islet. *Primates*, 6(1):1–30.

Kay, M. (1980). The proper place of men and machines in language translation. Technical Report CSL-80-11, Xerox Palo Alto Research Center, Palo Alto, CA. Available on `http://mt-archive.info/Kay-1980.pdf`.

Kay, M. (1997). Machine translation: The disappointing past and present. In in Chief), R. C. E., Mariani, J., Uszkoreit, H., Editor), G. B. V. M., Zaenen, A., Editor), A. Z. M., and Zue, V., editors, *Survey of the State of the Art in Human Language Technology*. Cambridge University Press and Giardini. Web Edition available at `http://www.dfki.de/~hansu/HLT-Survey.pdf`.

Kierkegaard, S. (1978). *Two ages: The Age of Revolution and The Present Age, A Literary Review*. Princeton University Press.

Kierkegaard, S. (1980). The concept of anxiety. Translated by Reidar Thomte.

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition*, 2(1):15–47.

King, M. (1984). When is the next alpac report due? In *ACL-22: Proceedings of the 10th International Conference on Computational Linguistics and 22nd annual meeting on Association for Computational Linguistics*, pages 352–353, Morristown, NJ, USA. Association for Computational Linguistics.

Kit, C. and Wong, T. M. (2008). Comparative Evaluation of Online Machine Translation Systems with Legal Texts. *Law Library Journal*, 100(2):299.

Knight, K. (1997a). Automating knowledge acquisition for machine translation. *AI Magazine*, 18(4):81.

Knight, K. (1997b). Automating knowledge acquisition for machine translation. *AI Magazine*, 18(4):81–96. Available on `http://www.aaai.org/ojs/index.php/aimagazine/article/-view/1323/1224`.

Knight, K. (1999). A Statistical MT Tutorial Workbook. Prepared in connection with the 1999 John Hopkins University summer workshop. Available on `http://www.isi.edu/natural-language/-mt/wkbk.rtf`.

Koehn, P. (2004a). *Pharaoh: a beam search decoder for phrase-based statistical machine translation models*. USC Information Sciences Institute. Available on `http://www.isi.edu/licensed-sw/-pharaoh/manual-v1.2.ps`.

Koehn, P. (2004b). Pharaoh: a beam search decoder for phrase-based statistical machine translation models. *Lecture Notes in Computer Science*, 3265:115–124.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Annual meeting-association for computational linguistics*, volume 45, page 2.

Koehn, P., Och, F., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, page 54. Available on `http://acl.ldc.upenn.edu/N/-N03/N03-1017.pdf`.

Kuhn, T. S. (1977). Second thoughts on paradigms. In *The Essential Tension: Selected Studies in Scientific Tradition and Change*. Chicago University Press.

Lakoff, G. (1990). *Women, fire, and dangerous things*. University of Chicago Press.

Lakoff, G. and Johnson, M. (1980). *Metaphors We Live By*. The University of Chicago Press, Chicago.

Langacker, R. W. (1999). *Grammar and conceptualization*. Mouton de Gruyter, Berlin.

Langacker, R. W. (2002). *Concept, Image, and Symbol: The Cognitive Basis of Grammar*, volume 1 of *Cognitive Linguistics Research*. Mouton de Gruyter, Berlin, 2nd edition.

Langacker, R. W. (2008). *Cognitive Grammar: A Basic Introduction*. Oxford University Press, Oxford.

Leacock, C. and Chodorow, M. (1998). Combining local context and WordNet sense similiarity for word sense disambiguation. In Fellbaum, C., editor, *WordNet: An electronic lexical database*. MIT press, Cambridge, MA.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26, New York, USA. Association for Computing Machinery. Available on `http://www.ims.uni-stuttgart.de/ schulte/Teaching/Korpus-Semantik-08/Referenzen/Senses/lesk-1986.pdf`.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304. Available on `http://web.cs.ualberta.ca/~tszhu/webmining/paper/ir/sim.pdf`.

Ljungberg, C. and Tabakowska, E., editors (2009). *Insistent Images*, volume 5 of *Iconicity in language and literature*. John Benjamins Publishing Company, Amsterdam. In press.

Locke, W. N. and Booth, A. D., editors (1955). *Machine Translation of Languages: Fourteen Essays*. Technology Press of the Massachusetts Institute of Technology, New York.

Maat, J. (2004). *Philosophical languages in the seventeenth century: Dalgarno, Wilkins, Leibniz*. Kluwer Academic Publishers, Dordrecht, the Netherlands.

MacDonald, N. (1954). Language translation by machine – a report of the first successful trial. *Computers and Automation*, 3(2):6–10. Available on `http://www.mt-archive.info/Macdonald-1954.pdf`.

Maeder, C., Fischer, O., and Herlofsky, W. J., editors (2005). *Outside-in — Inside-out*, volume 4 of *Iconicity in language and literature*. John Benjamins Publishing Company, Amsterdam.

Margolis, E. and Laurence, S. (1999). *Concepts: core readings*. The MIT Press.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist*, 44(12):1469–1481.

Medin, D. L. and Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology*, 20(2):158–190.

Meehan, J. R. (1977). TALE-SPIN, an interactive program that writes stories. In *In Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, pages 91–98, San Francisco, USA. IJCAI, Morgan Kaufmann Publishers. Available on `http://ijcai.org/Past%20Proceedings/IJCAI-77-VOL1/PDF/013.pdf`.

Meehan, J. R. (1981). Tale-spin. In Schank, R. C. and Riesbeck, C. K., editors, *Inside computer understanding: Five programs plus miniatures*, pages 197–226. Lawrence Erlbaum.

Merleau-Ponty, M. (1962). *Phenomenology of perception*. Burns & Oates.

Morland, D. V. (2002). Nutzlos, bien pratique, or muy util? business users speak out on the value of pure machine translation. Technical report, NCR Corporation, Dayton, Ohio. Available on `http://www.roi-learning.com/dvm/pubs/articles/tatc-24`.

Müller, W. G. and Fischer, O., editors (2003). *From Sign to Signing*, volume 3 of *Iconicity in language and literature*. John Benjamins Publishing Company, Amsterdam.

Nänny, M. and Fischer, O., editors (1999). *Form miming meaning*, volume 1 of *Iconicity in language and literature*. John Benjamins Publishing Company, Amsterdam.

Nänny, M. and Fischer, O., editors (2001). *The motivated sign*, volume 2 of *Iconicity in language and literature*. John Benjamins Publishing Company, Amsterdam.

Newman, S. M. (1957). Linguistic problems in mechanization of patent searching. Patent Office Research and Development Reports 9, Patent Office, U.S. Department of Commerce, Wahington, D.C. 1959 reprint available on `http://ipmall.info/hosted_resources/-patent-R&D-reports/Report_No_09.pdf` .

Nirenburg, S. (1989). Knowledge-Based Machine Translation. *Machine Translation*, 4(1):5–24.

NIST (2008). NIST 2008 Machine Translation Evaluation – (Open MT-08). Official evaluation results, National Institute of Standards and Technology, Multimodal Information Group, Gaithersburg, MD, USA. Available on `http://www.itl.nist.gov/iad/mig/tests/mt/2008/doc/mt-08_official_results_v0.html`.

Och, F. J., Tillmann, C., and Ney, H. (1999). Improved alignment models for statistical machine translation. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28.

Onyshkevich, B. and Nirenburg, S. (1995). A lexicon for knowledge-based MT. *Machine Translation*, 10(1):5–57.

Opdycke, J. B. (1949). *Mark My Words*. Harper & Brothers, New York, first edition.

Panov, D. Y. (1960). *Automatic Translation*. Pergamon Press, London. Translated from the Russian by R. Kisch.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). Wordnet::similarity – Measuring the Relatedness of Concepts. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1024–1025, Menlo Park, CA. Association for the Advancement of Artificial Intelligence, AAAI Press. Available on `http://acl.ldc.upenn.edu/N/N04/N04-3012.pdf`.

Perkins, R. L. (1985). *The Concept of anxiety*. Mercer University Press, Macon, Georgia.

Pierce, J. R., Carroll, J. B., Hamp, E. P., Hays, D. G., Hockett, C. F., Oettinger, A. G., and Perlis, A. (1966). Language and machines: Computers in translation and linguistics. Technical Report 1416, National Research Council, Wahington, D.C. Available on http://books.nap.edu/-openbook.php?record_id=9547.

Polanyi, M. (1966). *The tacit dimension*. Routledge & Kegan Paul, London.

Resnik, P. Using information content to evaluate semantic similarity in a taxonomy.

Richens, R. H. (1958). Interlingual machine translation. *The Computer Journal*, 1(3):144–147.

Richens, R. H. and Booth, A. D. (1955). Some methods of machanized translation. In Locke, W. N. and Booth, A. D., editors, *Machine Translation of Languages: Fourteen Essays*, pages 24–46. Technology Press of the Massachusetts Institute of Technology, New York.

Robinson, D. (1991). *The translator's turn*. Johns Hopkins University Press, Baltimore, Maryland.

Rorty, R. (1984). The historiography of philosophy: four genres. In Rorty, R., Schneewind, J. B., and Skinner, Q., editors, *Philosophy in History*, Ideas in context, pages 49–76. Cambridge University Press, Cambridge, UK.

Rorty, R. (1993). Wittgenstein, heidegger, and the reification of language. In Guignon, C. B., editor, *The Cambridge companion to Heidegger*, pages 337–357. Cambridge University Press, Cambridge.

Rumelhart, D. E. (1989). The Architecture of Mind: A Connectionist Approach. In Posner, M. I., editor, *Foundations of Cognitive Science*, pages 133–159. MIT Press, Cambridge, MA.

Russell, B. (1923). Vagueness. *Australasian Journal of Philosophy*, 1(2):84–92.

Ryle, G. (1949). *The Concept of Mind*. Hutchinson, London.

Sacks, H. (1995). *Lectures on conversation*. Blackwell Publishing, Malden, MA, USA.

Sadowski, P. (2001). The sound as an echo of the sense: The iconicity of English *gl-* words. In Fischer, O. and Nänny, M., editors, *The motivated sign*, volume 2 of *Iconicity in language and literature*, pages 69–88. John Benjamins Publishing Company, Amsterdam.

Sartre, J.-P. (2002). *Sketch for a Theory of the Emotions*. Routledge, London.

Sartre, J.-P. (2007). *Existentialism is a Humanism*. Yale University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656.

Shannon, C. E. (1949). Communication theory of secrecy systems. *Bell System Technical Journal*, 28:656–715.

Sheridan, P. (1955). Research in language translation on the ibm type 701. *IBM Technical Newsletter*, (9):5–24. Available on http://www.mt-archive.info/Sheridan-1955.pdf.

Silverstein, M. (1976). Hierarchy of features and ergativity. In *Grammatical categories in Australian languages*, pages 112–171. Australian Institute of Aboriginal Studies.

Simon, H. A. and Newell, A. (1958). Heuristic problem solving: The next advance in operations research. *Operations Research*, 6(1):1–10. Available on `http://or.journal.informs.org/cgi/reprint/6/1/1`.

Simons, D. J. and Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, 28:1059–1074. Available on `http://wexler.free.fr/library/files/simons%20(1999)%20gorillas%20in%20our%20midst.%20sustained%20inattentional%20blindness%20for%20dynamic%20events.pdf`.

Simpson, J., editor (2009). *Oxford English Dictionary Online*. Oxford University Press. `http://dictionary.oed.com/cgi/entry/00329286`.

Spinosa, C. and Dreyfus, H. L. (1996). Two kinds of antiessentialism and their consequences. *Critical Inquiry*, 22(4):735–763.

Stocklmayer, S. and Treagust, D. (1996). Images of electricity: how do novices and experts model electric current? *International Journal of Science Education*, 18(2):163–178.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Tucker, A. B. (1987). Current strategies in machine translation research and development. In Nirenburg, S., editor, *Machine Translation: Theoretical and methodological issues*, pages 22–41. Cambridge University Press, Cambridge, MA.

Turing, A. M. (1946). *Proposed Electronic Calculator*. London, UK. Available online on `http://www.turingarchive.org/viewer/?id=149&title=01`.

Turing, A. M. (1951). *Programmers' handbook for Manchester electronic computer Mark II*. Manchester, UK. Available on `http://turing.ecs.soton.ac.uk/browse.php/B/32`.

Uwe-Panther, C. and Radden, G. (1999). *Metonymy in language and thought*. Benjamins.

Vauquois, B. (1968). A survey of formal grammars and algorithms for recognition and transformation in machine translation. In *IFIP Congress-68*, pages 254–260, Edinburgh.

Vygotsky, L. S. (1986). Thought and language.

Weaver, W. (1955). Translation. In Locke, W. N. and Booth, A. D., editors, *Machine Translation of Languages: Fourteen Essays*, pages 15–23. Technology Press of the Massachusetts Institute of Technology, New York.

Weaver, W. and Shannon, C. E. (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Illinois.

White, J. S. and O'Connell, T. A. (1994). Evaluation in the ARPA machine translation program: 1993 methodology. In *Proceedings of the workshop on Human Language Technology*, pages 135–140, Morristown, NJ, USA. Association for Computational Linguistics.

Wilks, Y., Huang, X., and Fass, D. (1985). Syntax, preference, and right attachment. In Joshi, A. K., editor, *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, pages 779–784, Los Angeles, CA. Morgan Kaufmann.

Winograd, T. (1971). Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. Technical Report AI-TR-235, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts. Available on `http://dspace.-mit.edu/bitstream/handle/1721.1/7095/AITR-235.pdf?sequence=2`.

Winograd, T. (1991). Thinking machines: Can there be? are we? In Sheehan, J. J. and Sosna, M., editors, *The Boundaries of Humanity: Humans, Animals, Machines*, pages 198–223. University of California Press, Berkeley. The entire books is available on `http://citeseerx.ist.psu.edu/-viewdoc/download?doi=10.1.1.120.1623&rep=rep1&type=pdf` .

Winograd, T. and Flores, F. (2005). *Understanding computers and cognition: A new foundation for design*. Addison-Wesley, Reading, MA.

Wittgenstein, L. (1958a). *Philosophical Investigations*. Basil Blackwell, Oxford, second edition.

Wittgenstein, L. (1958b). *Preliminary studies for the "Philosophical Investigations" generally known as The Blue and Brown Books*. Harper & Row, New York.

Wittgenstein, L. (1969). *On Certainty*. Basil Blackwell, Oxford.

Wittgenstein, L. (1978). *Remarks on the Foundations of Mathematics*. Basil Blackwell, Oxford, third edition.

Wittgenstein, L. (1979). *Remarks on Frazer's Golden Bough*. The Brynmill Press Ltd., Retford.

Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics Morristown, NJ, USA. Available on `http://acl.ldc.upenn.edu/P/P94/-P94-1019.pdf`.

Yamamoto, M. (1999). *Animacy and reference: A cognitive approach to corpus linguistics*. John Benjamins Publishing Co., Amsterdam, the Netherlands.

Yngve, V. H. (1955). Sentence-for-sentence translation. *Mechanical Translation*, 2(2):29–37. Available on `http://www.mt-archive.info/MT-1955-Yngve.pdf`.

Yngve, V. H. (1957). A framework for syntactic translation. *Mechanical Translation*, 4(3):59–65. Available on `http://www.mt-archive.info/MT-1957-Yngve.pdf`.

Yngve, V. H. (2000). Early research at MIT: In search of adequate theory. In Hutchins, W. J., editor, *Eary Years in Machine Translation*, volume 97 of *Studies in the history of the language sciences*, pages 39–72. John Benjamins Publishing Co., Amsterdam, the Netherlands.

Yunker, J. (2008). The end of translation as we know it. *Multilingual*, 19(8):30–31.

Zaenen, A., Carletta, J., Garretson, G., Bresnan, J., Koontz-Garboden, A., Nikitina, T., O'Connor, M. C., and Wasow, T. (2004). Animacy encoding in english: Why and how. In Webber, B. and Byron, D. K., editors, *ACL 2004 Workshop on Discourse Annotation*, pages 118–125, Barcelona, Spain. Association for Computational Linguistics. Available on `http://www.aclweb.org/anthology-new/W/W04/W04-0216.pdf`.

Zahle, H. (2003). *Omsorg for retfærdighed: essays om retlig praksis*. Gyldendal, Copenhagen.