

## **Are some languages more complex than others? On text complexity and how to measure it**

*Iørn Korzen, Copenhagen Business School*

**Abstract:** In this paper, I discuss the concept of linguistic complexity, which has been high on the linguistic agenda during the last few decades (Merlini Barbaresi (ed.) 2003, Sampson et al. (eds.) 2009, Moretti 2018 and many others). I first cite the most important definitions of complexity proposed by different scholars, I then apply and compare particular elements of these definitions to linguistic phenomena found in two specific languages, Italian and Danish. I focus mainly on the number of propositions per sentence and on the degree of their subordination (as conveyed by verb implicitness and nominalisation), two manifestations of complexity that are numerically measurable and cross-linguistically comparable. I give both cross- and intralinguistic examples taken from comparable texts that exhibit differences in these kinds of complexity, and in this way I demonstrate that linguistic complexity is clearly linked to and dependent on the language type in question as well as the given uses and people. In the case of Italian, we might talk about a “language-internal multilingualism”. However, I conclude the paper by giving a positive answer to my question: Some languages are indeed more complex than others.

**Keywords:** Language complexity, text structure, text density, deverbalisation, intra- and cross-linguistic text comparison.

### **1. Introduction**

“Italian is a very complex and complicated language”. I am certain I am not the only Scandinavian language teacher who has heard such a lamentation or complaint from students, and probably not just about Italian, also more generally about the Romance languages. But is it true? Are some languages or language groups more complex than others? If so, by which parameters, and can it be measured and proved?

Some decades ago, the general answer from the (at that time relatively few) scholars who were investigating the concept of complexity was “no”. A well-known and often quoted statement is the following by Richard Hudson (1981: 2.2): “There is no evidence that normal human languages differ greatly in the complexity of their rules”, a statement that originally appeared in a paper in *Journal of Linguistics* with the eloquent title “83 things linguists can agree about”. The assumption, later baptized “ALEC” (“All Languages are Equally Complex”), implies that a relative simplicity in one respect in a language would entail a relative complexity in another respect in the same language (McWhorter 2001: 127-129), and was partly based on the supposition that all languages should generally be acquired equally readily by children as their first language (Hudson 1981: 2.5.i; Ferguson 1982: 61).

However, already a year later the axiom was questioned, and it was acknowledged that “like any other proposed general principle, it requires empirical verification” (Ferguson 1982: 62). It was later dismissed altogether, and it was recognised that some linguistic systems are indeed more difficult and more complex than others (Trudgill 2009: 99); cf. also McWhorter (2001), Moretti (2018) and most of the contributions to Sampson et al. (eds.) (2009). Much debate has also focused on whether language complexity is a universal constant or an evolving variable (Sampson 2009: 1), cf. e.g. many of the contributions to Sampson et al. (eds.) (2009), which also bears a very eloquent title: *Language Complexity as an Evolving Variable*. This question has been examined particularly in sociolinguistics, and mostly with a focus on the opposite concept of complexity, i.e. simplicity and simplification, cf. e.g. Fiorentino (2009) and Moretti (2018). Berruto (2012: 176) concludes – in agreement with my students – that instead of considering popular Italian as a simplified linguistic variety, one could and

should consider *standard* Italian as a particularly complex and in a certain sense “unnatural” linguistic variety due to its early literary and “elitist” standardization, cf. also Migliorini (1971).

But it should be specified which features or aspects of a given language we consider to be complex or difficult. It could be equally argued that when it comes e.g. to pronunciation and simple sentence structure, Italian is a fairly uncomplicated and easy language to learn. We shall return to the particularly challenging aspects of Italian in sections 3-4.

## 2. Definitions and examples of complexity

The notion of “complexity” came to linguistics from other sciences such as physics, anthropology and philosophy, and was basically defined as the quantity of information necessary to comprise and adequately understand and explain a system or set of elements (Merlini Barbaresi 2003: 24, 2005: 302). The same parameter is found in many definitions of linguistic complexity. Some scholars distinguish between “absolute or objective” and “relative, perceptual or agent-related” complexity, the former being intrinsic to the linguistic system proper and the latter the complexity experienced by language users as features that are costly or difficult to speakers, hearers and language learners (Masi 2003; Merlini Barbaresi 2004: sect. 1; Miestamo 2009: 81ff; Dahl 2009: 50-52; Fiorentino 2009: 282ff). On the particular difficulties that web texts create for the receiver, see Prada (2003).

In his sociolinguistic approach, Moretti (2018: 40-42) subdivides the systemic complexity in “spontaneous (or primary) complexity”, linked to the creation of the language in question, and a “second phase (or secondary) complexity”, caused e.g. by the diffusion of written language and normativisation of particularly formal varieties, a very evident characteristic of Italian.

Bisang (2009: 35, 37-38) suggests a distinction between “overt” and “hidden complexity”, where the former “forces the speaker to explicitly encode certain grammatical categories” (as e.g. in obligatory case marking in nouns or pronouns or mood marking in verbs), and the latter consists of either grammatical markers that have a wide range of meaning but are non-obligatory, or simple surface structures open to more than one interpretation, as e.g.,

(1) I saw the man in the park with the telescope. (Bisang 2009: 43),

where *in the park* can modify either *the man* or *I saw the man*, and *with the telescope* can modify either *the park* or *I saw the man in the park*. “Hidden complexity” always depends on pragmatic inference and makes quantitative analyses very difficult (Bisang 2009: 49).

Scholars generally agree with McWhorter (2001: 133) that “ranking any human language upon a complexity scale is a daunting task”. Deutscher (2009) states: “I argue that it is in fact impossible to define the notion of overall complexity in an objective, meaningful way. At best, the “overall complexity” of a language can be understood as “a vector (one-dimensional matrix) of separate values” (Deutscher 2009: 247), i.e. as the accumulation of different values that together bring the complexity to a higher or lower point in the same dimension. See also Bertuccelli (2003: 139), Maas (2009: 177), Nichols (2009: 120ff) and Moretti (2018: 37). Even the mere notion of textual complexity is a very complex one: “[T]extual complexity turns out to be the result of the cumulative effects of the interaction among different variables that belong to all levels of texture” (Masi 2003: 142).

The scholars that go into more specific definitions take their point of departure in the general approach to complexity of an object as “the amount of information needed to recreate or specify it” (Dahl 2009: 50). McWhorter (2001: 125) defines linguistic complexity as “the degree of overt signalling of various phonetic, morphological, syntactic, and semantic distinctions beyond communicative necessity”, and as cases of such “unnecessary overspecification” he mentions (McWhorter 2001: 161) gender marking, multiple past tenses, subjunctive marking and nominalised

intransitive verbs. In the same vein, Fiorentino (2009: 282-286) quotes as factors contributing to linguistic complexity a high number of subtypes of a given item, a high number of alternative variants of a (typically morphological) function, a high number of syntactic rules (and exceptions to the rules) to generate an output, i.e. more syntax than pragmatics, and less transparency in the relation form-function. For a similar definition, see Nichols (2009: 111-112).

Fiorentino is inspired e.g. by Ferguson (1982: 60), who gives a number of examples of complex linguistic structure, among which are: Larger lexicon in a given semantic domain, extensive inflectional systems, syntactically subordinate clauses and word order conditioned by syntax. A similar, but more detailed, overview can be found in Masi (2003: 140-141).

On the topic of second language acquisition (SLA), cf. also Ellis (2016). As particular difficulty factors in an SLA context, Ellis (2016: 342) lists “perceptual salience, semantic complexity, morphophonological regularity, syntactic category, and frequency”, of which perceptual salience plays a special role. The notion of salience is defined (Ellis 2016: 342) as “the ability of a stimulus to stand out from the rest”, and it “can be independently determined by physics and the environment”. Particularly important for perceptual salience are “the number of phones in the functor (phonetic substance), the presence/absence of a vowel in the surface form (syllabicity), and the total relative sonority of the functor”, and “linguistic forms of low psychophysical salience are more difficult both to perceive and to learn”. However, the constant of “learner-related complexity” is at best dubious since it will largely depend on the distance between the learners’ L1 and their L2 (Deutscher 2009: 247).

In an analysis of the differences between spoken and written language, Chafe (1985) introduces the concept of “idea units” that contain “all the information a speaker can handle in a single focus of consciousness” (Chafe 1985: 106). Such idea units are textualised as sentences in written language, and as “devices” for expanding the complexity of written idea units, Chafe mentions (Chafe 1985: 108-117) nominalisations, pre- and postposed present and past participles, participle clauses, dependent finite clauses and appositives. Non-finite verb forms are an important factor also in Merlini Barbaresi’s (2003: 40ff, 2004: sect. 6) analyses of text complexity in written recipes found in cookbooks.

Spoken idea units are generally shorter than written ones and more independent of each other. According to Chafe (Chafe 1985: 108), the average numbers in written and spoken English are 11 and 7 words per sentences, respectively.

### **3. The complexity of Italian (from a Danish viewpoint)**

As a teacher of Italian in Denmark for more than 40 years, my personal experience is that the most challenging complexity a Scandinavian student faces when learning Italian, and probably more generally a Romance language, is found at the text level. More precisely, some of the main challenges are indeed found exactly in the “complexity devices” outlined by Chafe and McWhorter, devices with which the Italians embellish their (especially written) idea units in their quest for “il bello stile”, the beautiful style.

This complexity is largely made possible by some typological phenomena and differences from Scandinavian; one very important one is the inventory of synthetic verb forms, which regarding Danish and Italian can be listed as illustrated in Table 1 – together with the linguistic distinctions expressed by the verb forms.

Table 1: Morphological differences between Danish and Italian verbs (synthetic forms)

Danish	Italian	Distinction
present, preterite	present, imperfetto, passato remoto, future	temporal/aspectual <sup>1</sup>
indicative, imperative	indicative, imperative, subjunctive, conditional	modal
infinite, present/past participle, nominalisation	infinite, present/past participle, gerund, nominalisation	rhetorical (foreground – background)

As Table 1 shows, Danish has fewer morphological possibilities than Italian to express temporal distinctions (no future form), no imperfetto/passato remoto to express an aspectual distinction, fewer modal distinctions (no subjunctive and conditional), and fewer rhetorical distinctions given the inexistence of the gerund and a much lower frequency of the other non-finite verb forms (see below). Since Danish functions quite well with such limitations, what Italian (together with the other Romance languages) has in addition to the Danish morphology, could, in McWhorter's (2001: 125) terminology, be considered as “distinctions beyond communicative necessity”, and hence as linguistic complexity.

The richer Romance inflection can be learned by Scandinavian language students and may even become a positive factor in text perception and decoding, as aptly expressed by Masi (2003: 141): “Regular morphological patterns that have been internalised by the reader ... may even help text decoding”. However, other text structural problems may arise. It could be hypothesised that the mentioned morphological richness could have particular consequences as it entails a higher tendency towards distinguishing, narratively and/or pragmatically, between propositions: aspectually, modally or rhetorically, as outlined in the right column of Table 1. Such distinguishing might logically imply a higher tendency to include more propositions in the same idea unit – precisely with the purpose to compare or distinguish between them. Since some propositions would be interpreted as aspectually, modally or rhetorically subordinated another proposition, the same analysis could imply a higher tendency towards subordination. Vice versa, less aspectual, modal or rhetorical differentiation, as in the Scandinavian languages, could imply less comparison, hence incorporation of fewer propositions in the same idea unit, and a higher tendency towards parataxis (Korzen 2017, 2018).

Can such hypotheses be tested and possibly proved?

Well, I would say at least partly. It is clear (from Table 1) that the Italian verb morphology is considerably richer and more complex than the Danish one. Furthermore, a comparison of the number of propositions per sentence and the degree of subordination of such propositions in comparable texts will – as we shall see in the following sections – show a clear tendency of a higher text structure complexity on both parameters in Italian texts than in Danish. In this context, comparable texts are understood as authentic texts produced independently of each other (hence not translations) but in equivalent situations, with an equivalent content and intended for equivalent receivers.

#### 4. The corpora

I conducted a number of quantitative and qualitative analyses on the basis of three different corpora of comparable Danish – Italian texts, argumentative, narrative, and expository texts respectively. The three corpora are:

- The *Europarl corpus*, consisting of political speeches held in the European Parliament (Koehn 2005, <http://statmt.org/europarl/>): argumentative texts.

<sup>1</sup> I here ignore the modal values that some of the mentioned verb forms can convey.

- The *Mr. Bean corpus*, 90 retellings of two Mr. Bean episodes produced by 18 Danish and 27 Italian university students (Skytte, Korzen, Polito, Strudsholm 1999, <http://blog.cbs.dk/mrbean-korpus/>): narrative texts. This corpus was produced in 1995 by a group of researchers from Copenhagen University and the Copenhagen Business School, including myself.
- The Danish project *SugarTexts - Telling the SugarStory in diverse languages*, led by our TYPOlex colleague Viktor Smith (Smith 2009, <http://www.sugartexts.dk/>): expository texts on the production of sugar from sugar beets.

The total number of words and the average number of words per sentence in these three corpora are cited in Table 2.

Table 2: Total number of words and words per sentence in three Danish-Italian corpora

		1.	2.	
		Total number of words	Average numbers	Difference
<b>Corpus Europarl</b>	<b>Danish</b>	14,737	33.4	53.9 %
	<b>Italian</b>	14,708	21.7	
<b>Corpus Mr. Bean</b>	<b>Danish</b>	7,262	22.8	14.0 %
	<b>Italian</b>	7,278	20.0	
<b>Sugar Texts</b>	<b>Danish</b>	4,851	24.8	67.6 %
	<b>Italian</b>	4,819	14.8	

In all three cases, the Danish and Italian corpora are practically of the same size regarding the number of words (column 1) – and fairly modest because all the following analyses are carried out manually. Moreover, the Italian sentences are longer, i.e. contain more words, than the Danish ones, the differences being most evident in the argumentative and expository texts.

Of course, reservations should be made when operating with calculations based on the unit “(graphic) word”. Some typological differences lead to a lower number of words in Danish (e.g. the definite article which in Danish is often enclitic and the many Danish nominal compounds that correspond to noun phrases N + preposition + N in Italian). On the other hand, other differences entail a lower number in Italian (e.g. verb + enclitic pronoun in Italian, non-existent in Danish, and the pro-drop phenomenon, i.e. the fact that the overt subject pronoun in Italian is only used for emphasis and otherwise omitted). However, most of the cross-linguistic differences mentioned in this and the following tables are of a size that in my opinion allows them to be used as indications of fundamental textual differences.

In itself, a simple number count of words per sentence may or may not be indicative of a high or low complexity. Even though Merlini Barbaresi (2004: sect. 2) claims that “Longer discourse units can be predicted to be more complex ... because every addition, during the progression of discourse, requires integration at a more general level”; Masi (2003: 141), on the other hand, states that “long structures and a great deal of linguistic information ... may even help text decoding”. Cf. also Fiorentino (2009: 309) who, in this matter, compares written and oral language:

il parlato, come è noto, produce più materiale linguistico per realizzare uno stesso significato se parliamo la versione scritta e orale di un discorso.

‘oral language, as is well-known, produces more linguistic material to achieve the same meaning if we compare the written and oral version of a speech’

#### 4.1. Propositions

In order to understand the purpose the “extra words” in Italian sentences serve, and thereby possibly get a more precise impression of text complexity phenomena and differences, I then counted the number of propositions per sentence, i.e. per “idea unit” in Chafe’s terminology. The propositions are the linguistic structures composed by a predicate + its connected arguments and related to a particular instance or circumstance, whereby a truth value can be ascribed to it (Herslund 1996: 37). The propositions can be said to constitute the “ideational skeleton” of a sentence (Herslund & Smith 2003: 113-114), i.e. in Halliday’s terminology (e.g. Halliday 1985: xiii) the basic elements of reference to the environment, the other two metafunctional components of linguistic communication in Halliday’s system being the interpersonal (the communication situation and its components) and the textual (the way in which the text is structured).

The propositions can be described as formulas  $P (A_1, A_2, A_3)$ , where  $P$  is the predicate and  $A_{1-3}$  the arguments. Applied to the two sentences in (2), an example taken from the Sugar Text corpus:

- (2) Når vognmanden kommer til sukkerfabrikken, kører han først op på en vægt. Her bliver lastbilen med roerne vejet. Når læsset er tørt af, bliver den tomme lastbil vejet igen. (DA1)  
[liter. transl.] ‘When the truck driver arrives at the sugar factory, he first drives on to a scale. Here the truck with the beets is weighed. When the load is emptied, the empty truck is weighed again.’

the propositions could be described in this way:

- (2)’ sentence 1: *komme (vognmanden, til sukkerfabrikken)*  
*køre (han, op på en vægt)*  
sentence 2: *veje (lastbilen)*  
sentence 3: *tømme af (læsset)*  
*veje (den tomme lastbil)*

As illustrated in (2)’, if the predicates are verbs, which is most often the case, these are typically cited in their infinitive form, the actual morphological realisation (tense/mood/aspect/diathesis/finiteness etc., as illustrated in Table 1) belonging to the metafunctional textual level in Halliday’s system. However, the description in (2)’ clearly illustrates a fairly simple text structure in (2), with only one or two propositions per sentence, which we shall return to below.

#### 4.2. Non-finite verbs

I also counted the number of non-finite realisations, i.e. implicit verbs (infinitives, participles, gerund), and nominalisations. These forms, which Chafe includes as “complexity devices”, are particularly interesting in that – deverbalised as they are in lacking modal, temporal and aspectual values – they unambiguously mark the proposition in question as a rhetorical satellite, i.e. a clause expressing a background situation with regard to a rhetorical nucleus (Matthiessen & Thompson 1988; Fox 1987; Rigotti 1993; Korzen 2007, 2009, 2017)<sup>2</sup>. In my analysis I therefore only include implicit and nominalised verb forms that function as predicates in propositions, and thus can be substituted by finite verb forms, see examples in (3)-(5) below.

These forms are also very interesting in connection with the notion of “text density” and the distinction between more or less “binding texts”. Text density is defined as the relation between the

<sup>2</sup> Nominalisations constitute an exception to that rule in that they may function as valency constituents, e.g. *L’arrivo di Luca non era previsto da nessuno*. ‘Luca’s arrival was not expected by anyone’, in which case the nucleus-satellite relation is much more complicated. For discussions on this topic, see e.g. Korzen (1999: 348-354; 2000: 86-88).

quantity of linguistic material of a text (span) and the information that the text (span) in question intends to express (Fabricius-Hansen 1996, 1998, 1999; Jansen 2003; Hansen-Schirra, Neumann & Steiner 2007; Korzen & Gylling 2017). (For binding/non-binding texts, see below.) Non-finite verb forms are perfect examples of dense text structures since they convey a textual content using less linguistic material than an equivalent finite text structure, which will typically contain a conjunction, a finite auxiliary verb, and a subject (Korzen 2014, 2015). Cf. the following two (very typical) Italian non-finite structures in (a), a past participle and a gerund respectively, compared with the finite structures in (b), more typical in Danish, as we shall see below:

- (3) a. *Arrivato tardi*, Luca ha perso l'inizio del film. '*Having arrived late*, Luca missed the beginning of the film'.
- b. *Eftersom Luca var ankommet sent*, gik han glip af begyndelsen af filmen. '*Since Luca had arrived late*, he missed the beginning of the film'.
- (4) a. *Arrivando tardi*, perderai l'inizio del film. '*By arriving late*, you'll miss the beginning of the film'.
- b. *Hvis du kommer for sent*, går du glip af begyndelsen af filmen. '*If you arrive late*, you'll miss the beginning of the film'.

In all cases, the main clause *Luca missed the beginning of the film* / *You'll miss the beginning of the film* constitutes the nucleus part of the short texts, and the implicit structure / subordinate clause conveys the satellite expressing either cause (3) or condition (4). In (5), the nominalised satellite expresses time:

- (5) a. *All'arrivo di Luca*, siamo andati al cinema. '*At Luca's arrival*, we went to the cinema'.
- b. *Da Luca var ankommet*, gik vi i biografen. '*When Luca had arrived*, we went to the cinema'.

At the same time, the non-finite structures are less binding since they leave a precise semantic interpretation (e.g. cause, condition or time) up to the receiver to a much higher degree than a finite – and more explicit – structure. Cf. Sabatini (1999) who proposes a text typology based on the “communicative pact” between the speaker and the receivers including the degree with which the speaker binds the receivers in their interpretation of the text. A higher linguistic density requires a greater interpretative commitment by the receiver, in other words: it is more complex.

#### 4.3. Count results and analyses

The number of propositions per sentence and the percentage of non-finite textualisations in the mentioned corpora turned out as is shown in Table 3.

Table 3: Propositions per sentence and percentage of non-finite textualisations in three corpora

		1. Propositions per sentence		2. Implicit verb forms	3. Nominalisations
		Average numbers	Difference		
<b>Corpus Europarl</b>	<b>Danish</b>	2.15	55.8 %	13.4 %	1.8 %
	<b>Italian</b>	3.35		25.5 %	4.3 %
<b>Corpus Mr. Bean</b>	<b>Danish</b>	3.28	32.6 %	15.7 %	2.4 %
	<b>Italian</b>	4.35		33.7 %	5.9 %
<b>Sugar Texts</b>	<b>Danish</b>	2.54	82.3 %	8.9 %	11.5 %
	<b>Italian</b>	4.63		26.7 %	25.6 %

In all three cases, the number of propositions per sentence is considerably higher in the Italian texts (column 1), and both implicit and nominalised verbs (columns 2-3) are more than twice as frequent in Italian as in Danish (except the implicit verbs in the Europarl Corpus, which are however almost twice as frequent). Clearly these two accounts show a much higher text complexity in Italian than in Danish.

Whereas the Europarl and Bean corpora are fairly homogeneous, i.e. the individual texts vary relatively little from the average numbers cited in Table 3, the Sugar Text corpus displays considerable intralinguistic differences regarding these values, differences that are illustrated in the following Danish and Italian examples, of which I repeat ex. (2) from section 4.1:

- (2) Når vognmanden kommer til sukkerfabrikken, kører han først op på en vægt. Her bliver lastbilen med roerne vejet. Når læsset er tømt af, bliver den tomme lastbil vejet igen. (DA1) [liter. transl.] ‘When the truck driver arrives at the sugar factory, he first drives on to a scale. Here the truck with the beets is weighed. When the load is emptied, the empty truck is weighed again.’
- (6) La barbabietola immagazzina nella sua radice lo zucchero che fabbrica. Una volta raccolta essa viene trasportata velocemente allo zuccherificio. ...  
Lo zuccherificio è una fabbrica molto moderna: tutto il lavoro viene svolto da macchine, dall’arrivo delle barbabietole alla partenza dello zucchero. (IT10)  
[liter. transl.] ‘The sugar beet stores in its root the sugar that it produces. Once harvested it is quickly transported to the sugar factory.  
The sugar factory is a very modern factory: all the work is done by machines, from the arrival of the beets to the departure of the sugar.’
- (7) Ved fremstilling af sukker fra sukkerrør og sukkerroer bliver plantematerialet vasket og findelt, sukkeret ekstraheret og den resulterende sukkersaft renses og inddampes, og endelig bliver sukkeret isoleret og tørret. (DA10)  
‘In the production of sugar from sugar cane and sugar beet, the plant material is washed and finely divided, the sugar extracted, and the resulting sugar juice purified and evaporated, and finally the sugar is isolated and dried.’
- (8) Dopo averlo grossolanamente filtrato attraverso depolpatori che trattengono le particelle di fettucce tenute in sospensione, il sugo viene depurato per aggiunta di calce, spesso sotto forma di latte di calce (*defecazione*), che agisce trasformando i sali di calcio insolubili, gli acidi liberi

presenti e i loro sali alcalini, facendo variare il pH del mezzo e coagulando parte dei colloidî.  
(IT8)

[liter. transl.] ‘After having been filtered roughly through pulpers that retain the particles of cossettes held in suspension, the moisture is purified by the addition of lime, often in the form of milk of lime (*defecation*), which acts by transforming the insoluble calcium salts, the free acids present and their alkaline salts, by varying the pH of the substance and coagulating part of the collodion.’

In (2), we have three short sentences with a very simple syntax, one or two propositions per sentence, see (2)’ above, and only finite verb forms, that is: only one level of subordination. Also in (6), the sentences are short and syntactically simple; the text contains one implicit verb (*raccolta* ‘harvested’) and two nominalisations (*arrivo, partenza* ‘arrival, departure’). In (7), the sentence is longer, there are several propositions per sentence, however most of them coordinated, and one implicit verb form, *resulterende* ‘resulting’, and one nominalisation, *fremstilling* ‘production’. The Italian example in (8) displays the longest sentence, the most complex syntax and the highest number of implicit verb forms, *averlo filtrato, tenute, trasformando, facendo variare, coagulando* ‘having filtered it, held, transforming, varying, coagulating’, and nominalisations, *sospensione, aggiunta, defecazione* ‘suspension, addition, defecation’.

Thus, the differences regard precisely the values quoted in Table 3, and precisely the texts cited in (2), (6)-(8) differ most from the average values. Texts (2) and (6) are well below the average values, whereas texts (7) and (8) are above (except for the implicit verb forms in text (8)), cf. Table 4:

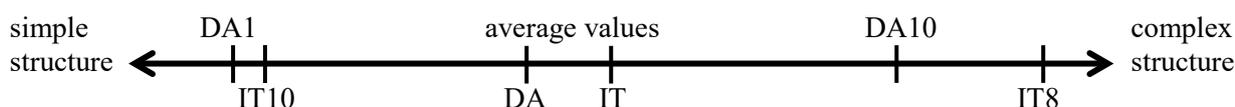
Table 4: Propositions per sentence and percentage of non-finite textualisations in four Sugar Texts

Texts	1. Propositions per sentence	2. Implicit verb forms	3. Nominal- isations
(2) DA1	1.66	0	0
(7) DA10	3.56	12.5 %	18.8 %
(6) IT10	2.31	16.7 %	6.7 %
(8) IT8	8.80	20.5 %	38.6 %

One very good reason for these differences is the fact that texts (2) and (7) are actually NOT perfectly comparable with (6) and (8); we are here dealing with different receiver types, in other words with different uses and different people. The Danish text (2), DA1, is an illustrated description, *Sukker*, ‘Sugar’, by Nanna Gyldenkærne (København, Mallings 1984), intended for the first five years of school. Similarly, text (6), IT10, consists of two short chapters from the book *Il tuo primo libro della fattoria*, ‘Your first book about the farm’ by Emilia Beaumont (Milano, Larus 2004) for children between 5 and 8 years of age. On the other hand, texts (7) and (8), DA10 and IT8, are both encyclopaedic: the entry *Sukker - Fremstilling* (‘Sugar - Production’) of *Den Store Danske Encyklopædi* (Copenhagen 2017), and the entry *Produzione dello Zucchero* (‘Sugar production’) of the *Italiano Dizionario Enciclopedico Italiano* (Rome 1961).

If we imagine a relative complexity scale, based on the two parameters suggested here, with the most simple and the most complex text structure at each end, the figure could look something like this:

Figure 1: The Sugar Texts complexity based on the two suggested parameters



In all three cases: the most simple texts, the most complex texts, and the average values, the Italian texts have higher values than the Danish ones, and the Italian intralinguistic differences are greater than the Danish ones except for the implicit verb forms. It is particularly interesting that the Italian values are higher than the Danish ones in the two simple texts since the Danish text, DA1, is intended (also) for older children than the target group of the Italian book, IT10.

The greater intralinguistic differences in the Italian texts, which are manifestations of the different uses and different people in play, one of the leitmotifs of this issue of *Globe*, make it tempting to talk about a kind of “language-internal multilingualism” in Italian, which can be seen as the background also for the phenomena described by Gargiulo (this issue). If such a concept is reasonable, this will no doubt be the case in many languages, but in some more than in others (as we have just seen), and precise definitions and criteria should be proposed and pursued. However, such a topic exceeds the framework of the present paper, and I shall leave it for (hopefully) later discussions and debates.

## 5. Conclusions and discussion

As many scholars have stated, it is probably impossible to determine and define such a notion as the “overall complexity” of a language; at best, it can be seen as the accumulation of many separate values, as Deutscher (2009: 247) expresses it. The two parameters suggested in this paper, the number of propositions per sentence and their degree of subordination, are of course by no means the only indicators of the complexity of a text. However, my personal experience as a teacher of Italian in Denmark for over 40 years tells me that they may very well be the two most difficult Italian text structure phenomena for a Dane to master perfectly. From a pedagogical point of view, I can confirm that they are indeed relevant, and I think the above observations show that they can function as useful scales by which we can measure two elements of the set of values that contribute to determining the complexity of a text. In my view, the data quoted in Tables 3-4 confirm the usefulness of the two parameters in both intra- and cross-linguistic comparisons, and from a text structure perspective, I am much inclined to agree with my students that, *ceteris paribus*, i.e. talking about the same uses and people, Italian (and generally Romance) text structure does tend to be more complex than Scandinavian structure.

I have, at least partially, linked the two text structure phenomena and the cross-linguistic differences to typological differences in verb morphology, cf. Table 1. It should be emphasised that these typological differences are absolutely not the only reason for a predisposition to a more complex text structure in Italian than in Danish. Substantial cultural-historical differences, such as the great importance given to Italian text composition and text structure and the high value set upon written and literary Italian, play a very important role as well (D’Achille 2003). The ability to express oneself in an appropriate manner according to the communication situation is of much higher socio-cultural importance in Italy than in Denmark, and linguistic and communicative competences are very highly regarded in the Italian school and university system (Korzen 2003, 2019).

In some cases, Danish has other ways of expressing the mentioned – or other – nuances, for instance modal distinctions such as those conveyed by the modal particles *jo*, *nu*, *da*, *vel*, *skam*, *dog*. There is little doubt that these particles constitute a big problem for learners of Danish L2, see also Lundquist (this issue), but I do not consider their ability to instantiate a (potential) opposition to a parallel proposition, e.g. without a particle, to be as strong as that of modal verb forms like the Italian

subjunctive and conditional. Such verb forms more overtly express what Reinhart (1984: 802) calls “alternative modes of events”: “Modal propositions (including ‘irrealis’ statements of alternative modes of events) ... are background. Such propositions function as clues for the understanding of the foreground by comparing its events to alternative modes of development”.

Danish can express similar irrealis statements e.g. by means of the modal verb *ville* ‘will’ in the preterite or past perfect. In such cases, we are dealing with “alternative modes of events” also in this language, but as a “hidden complexity” in Bisang’s (2009) terminology (cf. section 2 above), i.e. with what may look simple on the surface but can be based on a more complex background of different potential inferences.

## References

- Berruto, Gaetano (2012). ‘L’italiano popolare e la semplificazione linguistica’. In Gaetano Berruto, *Saggi di Sociolinguistica e Linguistica* a cura di Giuliano Bernini et al. Alessandria: Edizioni dell’Orso. 141-181. [First published in *Vox Romanica*, 42 (1983): 38-79].
- Bertuccelli, Marcella (2003). ‘Cognitive complexity and the lexicon’. In Lavinia Merlini Barbaresi (ed.), *Complexity in Language and Text*. Pisa: Edizioni Plus. 67-115.
- Bisang, Walter (2009). ‘On the evolution of complexity: sometimes less is more in East and mainland Southeast Asia’. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press. 34-49.
- Chafe, Wallace L. (1985). ‘Linguistic differences produced by differences between speaking and writing’. In David R. Olson, Nancy Torrance & Angela Hildyard (eds.), *Literacy, Language and Learning. The Nature and Consequences of Reading and Writing*. Cambridge: Cambridge University Press. 105-123.
- D’Achille, Paolo (2003). *L’Italiano Contemporaneo*. Bologna: il Mulino.
- Dahl, Östen (2009). ‘Testing the assumption of complexity invariance: the case of Elfdalian and Swedish’. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press. 50-63.
- Deutscher, Guy (2009). ‘“Overall complexity”: a wild goose chase?’ In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press. 243-251.
- Ellis, Nick C. (2016). ‘Salience, cognition, language complexity, and complex adaptive systems’. *Studies in Second Language Acquisition*, 38: 341-351. doi:10.1017/S027226311600005X.
- Fabricius-Hansen, Cathrine (1996). ‘Informational density - a problem for translation and translation theory’. *Linguistics*, 34: 521-565.
- Fabricius-Hansen, Cathrine (1998). ‘Information density and translation, with special reference to German – Norwegian – English’. In Stig Johansson & Signe Oksefjell (eds.), *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. Amsterdam: Rodopi. 197-234.
- Fabricius-Hansen, Cathrine (1999). ‘Information packaging and translation. Aspects of translational sentence splitting (German – English/Norwegian)’. In Monika Doherty (ed.), *Sprachspezifische Aspekte der Informationsverteilung*. Berlin: Akademie-Verlag. 175-213.
- Ferguson, Charles A. (1982). ‘Simplified registers and linguistic theory’. In Loraine K. Obler & Lise Menn (eds.), *Exceptional Language and Linguistics*. New York et al.: Academic Press. 49-68.
- Fiorentino, Giuliana (2009). ‘Complessità linguistica e variazione sintattica’. *Studi Italiani di Linguistica Teorica e Applicata*, XXXVIII(2): 281-312.
- Fox, Barbara A. (1987). *Discourse Structure and Anaphora. Written and Conversational English*. Cambridge: Cambridge University Press.
- Gargiulo, Marco (2021). ‘Language conflict, glottophagy and camouflage in the Italian cinematic city’. *Globe: A Journal of Language, Culture and Communication*, 12: 117-130.
- Halliday, Michael A. K. (1985). *An Introduction to Functional Grammar*. London: Arnold.

- Hansen-Schirra, Silvia, Stella Neumann & Erich Steiner (2007). 'Cohesive explicitness and explicitation in an English-German translation corpus'. In *Languages in Contrast*, 7(2): 241-265.
- Herslund, Michael (ed.) (1996). *Det Franske Sprog*. København: Handelshøjskolen i København.
- Herslund, Michael & Viktor Smith (2003). 'Semantik'. In Michael Herslund & Bente Lihn Jensen (eds.), *Sprog og Sprogbeskrivelse*. København: Samfundslitteratur. 83-125.
- Hudson, Richard (1981). *Some issues on which linguists can agree*. URL: <https://www.llas.ac.uk/resources/gpg/135.html>. Retrieved 16 August 2019. (The article first appeared in *Journal of Linguistics*, 17: 333-344 (1981) under the title '83 things linguists can agree about').
- Jansen, Hanne (2003). *Densità Informativa: Tre Parametri Linguistico-Testuali. Uno Studio Contrastivo Inter- ed Intralinguistico*. Copenhagen: Museum Tusulanum Press.
- Koehn, Philipp (2005). 'Europarl: A parallel corpus for statistical machine translation'. *Conference Proceedings: The Tenth Machine Translation Summit*. Thailand: Phuket. 79-86.
- Korzen, Iørn (1999). 'Tekststruktur og anafortypologi. (Struttura testuale e tipologia anaforica)'. In Gunver Skytte, Iørn Korzen, Paola Polito & Erling Strudsholm (eds.) (1999), *Tekststrukturering på Italiensk og Dansk. Resultater af en Komparativ Undersøgelse / Strutturazione Testuale in Italiano e Danese. Risultati di una Indagine Comparativa*. København: Museum Tusulanum Press. 331-418.
- Korzen, Iørn (2000). 'Tekstsekvenser: struktur og opbygning.' In Gunver Skytte & Iørn Korzen, *Italiensk–dansk Sprogbrug i Komparativt Perspektiv. Reference, Konnexion og Diskursmarkering*. København: Samfundslitteratur. 67-99.
- Korzen, Iørn (2003). 'Hierarchy vs. linearity. Some considerations on the relation between context and text with evidence from Italian and Danish'. In Irene Baron (ed.), *Language and Culture, Copenhagen Studies in Language*, 29: 97-109.
- Korzen, Iørn (2007). 'Linguistic typology, text structure and appositions'. In Iørn Korzen, Marie Lambert & Hélène Vassiliadou (eds.), *Langues d'Europe, l'Europe des langues. Croisements Linguistiques. Scolia*, 22: 21-42.
- Korzen, Iørn (2009). 'Struttura testuale e anafora evolutiva: tipologia romanza e tipologia germanica.' In Iørn Korzen & Cristina Lavinio (eds.), *Lingue, Culture e Testi Istituzionali*. Firenze: Franco Cesati. 33-60.
- Korzen, Iørn (2014). 'Struttura testuale e anafora nella traduzione del discorso politico: un'indagine tipologico-comparativa'. In Enrico Garavelli & Elina Suomela-Härmä (eds.), *Dal Manoscritto al Web: Canali e Modalità di Trasmissione dell'Italiano. Tecniche, Materiali e Usi nella Storia della Lingua: Atti del XII Congresso SILFI*. Firenze: Franco Cesati. 391-400.
- Korzen, Iørn (2015). 'Frase complesse e complessità frasale: il discorso politico in un'ottica tipologico-comparativa'. In Carla Bruno, Simone Casini, Francesca Gallina & Raymond Siebetecheu (eds.), *Plurilinguismo/Sintassi. Atti del XLVI Congresso Internazionale SLI*. Roma: Bulzoni. 625-642.
- Korzen, Iørn (2017). 'Struttura testuale e interpretazione nella traduzione da una lingua scandinava all'italiano'. In Vera Nigrisoli Wärnhjelm, Alessandro Aresti, Gianluca Colella & Marco Gargiulo (eds.), *Edito, Inedito, Riedito. Saggi dall'XI Congresso degli Italianisti Scandinavi*. Pisa: Pisa University Press 2017. 59-73.
- Korzen, Iørn (2018). 'L'italiano: una lingua esocentrica. Osservazioni lessicali e testuali in un'ottica tipologico-comparativa'. In Iørn Korzen (ed.), *La Linguistica Italiana nei Paesi Nordici. Studi Italiani di Linguistica Teorica e Applicata*, XLVII(1): 15-36.
- Korzen, Iørn (2019). 'Anaphors and text structure in Romance and Germanic languages: Typologies in comparison'. In Irène Baron, Louis Begioni, Michael Herslund & Alvaro Rocchetti (eds.), *Le Lexique et ses Implications : entre Typologie, Cognition et Culture. Langages*, 214(2): 75-

90.

- Korzen, Iørn & Morten Gylling (2017). 'Text structure in a contrastive and translational perspective: On information density and clause linkage in Italian and Danish'. In Oliver Czulo & Silvia Hansen-Schirra (eds.), *Crossroads between Contrastive Linguistics, Translation Studies and Machine Translation*. Berlin: Language Science Press. 31-64. <https://zenodo.org/record/1019687>.
- Lundquist, Lita (2021). 'Humour socialisation. Why the Danes are not as funny as they think they are'. *Globe: A Journal of Language, Culture and Communication*, 12: 32-47.
- Maas, Utz (2009). 'Orality versus literacy as a dimension of complexity'. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press. 164-177.
- Masi, Silvia (2003). 'The literature on complexity'. In Lavinia Merlini Barbaresi (ed.), *Complexity in Language and Text*. Pisa: Edizioni Plus. 117-145.
- Matthiessen, Christian & Sandra A. Thompson (1988). 'The structure of discourse and 'subordination''. In John Haiman & Sandra A. Thompson (eds.), *Clause Combining in Grammar and Discourse*. Amsterdam/Philadelphia: John Benjamins. 275-329.
- McWhorter, John H. (2001). 'The world's simplest grammars are creole grammars'. *Linguistic Typology*, 5: 125-166.
- Merlini Barbaresi, Lavinia (2003). 'Towards a theory of text complexity'. In Lavinia Merlini Barbaresi (ed.), *Complexity in Language and Text*. Pisa: Edizioni Plus. 23-66.
- Merlini Barbaresi, Lavinia (2004). 'Levels of text complexity'. In Piet van Sterkenburg (ed.), *Linguistics Today – Facing a Greater Challenge*. Amsterdam/Philadelphia: John Benjamins. CD-Rom.
- Merlini Barbaresi, Lavinia (2005). 'Il discorso economico/argomentativo: marcatezza e complessità della previsione'. In Leandro Schena, Chiara Preite & Sara Vecchiato (eds.), *Gli Insegnamenti Linguistici nel Nuovo Ordinamento: Lauree Triennali e Specialistiche dell'Area Economico-Giuridica*. Milano: Egea. 301-324.
- Miestamo, Matti (2009). 'Implicational hierarchies and grammatical complexity'. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press. 80-97.
- Migliorini, Bruno (1971<sup>4</sup>). *Storia della Lingua Italiana*. Firenze: Sansoni.
- Moretti, Bruno (2018). 'Che cosa ha da dire la sociolinguistica sul tema della complessità delle lingue'. *Rivista Italiana di Dialettologia*, 42: 35-52.
- Nichols, Johanna (2009). 'Linguistic complexity: a comprehensive definition and survey'. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press. 110-125.
- Prada, Massimo (2003). 'Lingua e web'. In Ilaria Bonomi, Andrea Masini & Silvia Morgana (eds.), *La Lingua Italiana e i Mass Media*. Roma: Carocci. 249-289.
- Reinhart, Tanya (1984). 'Principles of gestalt perception in the temporal organization of narrative texts'. *Linguistics*, 22(6): 779-809.
- Rigotti, Eddo (1993). 'La sequenza testuale: definizione e procedimenti di analisi con esemplificazioni in lingue diverse'. *L'analisi Linguistica e Letteraria*, 1: 43-148.
- Sabatini, Francesco (1999). "'Rigidità-esplicitzza" vs "elasticità-implicitzza": possibili parametri massimi per una tipologia dei testi'. In Gunver Skytte & Francesco Sabatini (eds.), *Linguistica Testuale Comparativa*. Copenhagen: Museum Tusulanum Press. 141-172.
- Sampson, Geoffrey (2009). 'A linguistic axiom challenged'. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press. 1-18.
- Skytte, Gunver, Iørn Korzen, Paola Polito & Erling Strudsholm (eds.) (1999). *Tekststrukturering på*

*Italiensk og Dansk. Resultater af en Komparativ Undersøgelse / Strutturazione Testuale in Italiano e Danese. Risultati di una Indagine Comparativa.* Copenhagen: Museum Tusulanum Press.

Smith, Viktor (2009). 'Telling the SugarStory in seven Indo-European languages. What may and what must be conveyed?' In Iørn Korzen & Cristina Lavinio (eds.), *Lingue, Culture e Testi Istituzionali*. Firenze: Franco Cesati. 61-76.

Trudgill, Peter (2009). 'Sociolinguistic typology and complexification'. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press. 98-109.

### Corpus texts

DA1: Gyldenkærne, Nanna (1984). *Sukker*. Copenhagen: Mallings.

DA10: 'Sukker – fremstilling' ('Sugar – production'). *Den Store Danske Encyklopædi*. Copenhagen 2017, <https://denstoredanske.lex.dk/sukker>.

IT8: Produzione dello zucchero ('Sugar production'), *Dizionario Enciclopedico Italiano* (1961). Vol. XII. Rome: Istituto dell'Enciclopedia Italiana. 1018-1019.

IT10: Beaumonte, Emilia (2004). *Il Tuo Primo Libro della Fattoria*. Milano: Larus.

--

Iørn Korzen

Copenhagen Business School

[ik.msc@cbs.dk](mailto:ik.msc@cbs.dk)