Journal of Universal Language 8
September 2007, 7-38

Redundancy Elimination: The Case of Artificial Languages

Isabella Chiari
University of Rome La Sapienza

Abstract

This paper discusses how artificial languages deal with redundancy at the theoretical level of planning and in their actual textual manifestations. After a brief introduction to the notion of redundancy in linguistics and information theory, we propose a new definition, under the light of which we observe different expressions in artificial languages, mostly Esperanto. Redundancy is strictly connected to predictability, uneven frequencies, functional asymmetries and syntagmatic constraints. In natural languages it represents a constitutive principle of communication, being present in all semiotic systems, including animal communication. What happens in artificial languages? Is redundancy expression a topic of discussion in language planning? How does it manifest in languages like Esperanto? A review of some of the major issues in language design is presented, underlining which features of redundancy can be eliminated and which are constitutive of any language, whether natural or artificial. Redundancy of graphemic systems, distributional redundancy of phonotactics, allomorphy, agreement and government as expressions of functional redundancy, and word order distributional properties will be considered as key manifestations of redundancy, observing how they are portrayed in planned languages.

Keywords: redundancy, information, natural languages, artificial languages, international languages, language construction

1. Introduction

Among the features of natural languages and other semiotic systems lies the property of redundancy (Chiari 2002), which characterizes both grammatical systems and their concrete realizations in spoken and written texts. It has been often claimed that artificial languages should avoid redundancies at different levels, thus aiming at maximum economy and optimality. Redundancy joins ambiguity, opacity, polysemy and synonymy among negative properties typical of historical natural languages which are focused on in order to be avoided or limited in planned artificial and international languages. Along with an aim at ease of learning and cultural neutrality (Sapir 1925), most artificial language planning involves some reflection on practical requirements such as the simplification of grammatical and lexical structures that are considered superfluous or unconstructive to reach maximum “simplicity, regularity, logic, richness, and creativeness” (Sapir 1925: 65). Even though discussion can be raised on actual concrete and undisputable definitions of the latter concepts, it is clear that a long debate about general design tendencies for artificial languages took place in the past and is in many aspects still alive.

In the following paragraphs a clarification of some aspects concerning the notion of redundancy in languages will be proposed in order to observe which features of redundancy are present in artificial languages (namely in Esperanto and Ido), which ones are avoidable, and which ones are constitutive of any language, whether natural or artificial.

2. Redundancy: A Definition

The concept of redundancy is often associated with the idea of superfluity, overabundance and surplus, being thus perceived negatively in its general meaning. The notion [1] was present in linguistic debates since the antiquity, with a general meaning of abundance and also develops an extended meaning in the field of rhetorics, where it is connected to pleonasm, accumulation, repetition.[2] According to Cohen (1963: 61) the form specializes in its rhetorical domain in the 16th century.

In language studies it appears sparsely without a proper definition until the formalization introduced by Shannon's theory of information (Shannon & Weaver 1949: 61; Shannon 1951). Redundancy in this framework is seen as the counterpart of information, described as the degree of uncertainty of a definite event to occur. The measure of information, being identified with the formula used in physics to describe *entropy* (H) is thus (Shannon & Weaver 1949):

(1) $H = - K \Sigma p_i \log p_i$

[1] The term *redundancy* derives from Latin *unda* (from Indoeuropean **ud-* 'water', as in Sanskrit *udnás,*), associated with the prefix *red-* meaning 'to overflow'), which produces a large number of forms: *abundare, inundare, superabundare, superundare* and *redundare*. In English *redundant* and *redundance* are probably loaned from middle French.

[2] The idea of redundancy in its rhetorical sense is not only common to ancient reflection on the qualities of text (brevity, being the compromise against excess and poverty - *quantum opus est, quantum satis est*, but is also present, disguised in Grice's maxim of quantity and manner (1975): Make your contribution to the conversation as informative as necessary; do not make your contribution to the conversation more informative than necessary; be brief (avoid unnecessary wordiness).

where K is a positive constant depending on the measurement unit, and *p* is the probability of an event to occur. The general interpretation of the formula is that the amount of information of a source equals minus the sum of the probabilities of each sign available at the source, each probability times its own logarithm. An increase in entropy means a decrease in order. The minimum value for H being 0 when all probabilities except one equal to zero, thus making the actual event completely predictable. The maximum value for H being *log n* when all probabilities are even (*1/n*), thus representing a highest of uncertainty on the actual event that will occur. The concept of *relative entropy* is further introduced by Shannon to give account of the intermediate cases and is represented by the following formula:

(2) $H / \log n$

Redundancy is defined as "one minus relative entropy" (Shannon & Weaver 1949: 61), coming to measure the non-economy of the communication system and having a fundamental role in safeguarding transmission over noisy channels and thus being a *security device.*[3] In Weaver's words "This is the fraction of the structure of the message which is determined not by the free choice of the sender, but rather by the accepted statistical rules governing the use of the symbols in question. It is sensibly called redundancy, for this fraction of the message is in fact redundant in something close to the ordinary sense; that is to say, this fraction of the message is unnecessary (and hence repetitive or redundant) in the sense that if it were missing the message would still be essentially

[3] A complete account of the role of redundancy in languages, and specifically in information theory, with the illustration of the main methods proposed to measure redundancy can be found in Chiari (2002: 35-112).

complete, or at least could be completed." Thus redundancy is connected with the predictability of certain elements of the code in a message, being a consequence of uneven frequency of occurrence and of the presence of constraints (being in Gustav Herdan's terms an *index of structure*). Therefore redundancy can either be be due to repetition either to predictability imposed for example by a constraint (Campbell 1982: 68). The connection between the notion of redundancy and the statistical properties of language is explicitly stated by Shannon (1951: 50): "redundancy, on the other hand, measures the amount of constraint imposed on a text in the language due to its statistical structure."[4]

After the application of information theory to linguistics (though mainly on the orthographic version of written texts), the concept of redundancy becomes familiar to many areas of the language sciences, from phonology to morphosyntax, mainly thanks to the critical contributions of Roman Jakobson and George A. Miller. A new theoretical problem arises: the need of a specific definition that can take into account the complexity and variety of manifestations of redundancy that characterize language at each level of analysis and as a whole.[5] A few proposals have been made, among which

[4] Within the framework of information theory many measures of redundancy have been proposed and applied to different languages, mainly to their orthographic systems. Experimental methods have been compared to mathematical formulas in order to approximate to a comparable evalutation of the concept: experimental/ predictive methods, cloze procedures, constraint coefficients, revisions of Zipf's law, the gambling estimates method, the measure of the entropy of roots, relative efficiency, etc. For a more detailed analysis of these methods see Chiari (2002: 50-64) and extensively in my Ph.D. dissertation (Chiari, 2000).

[5] The terminological and conceptual confusion dealing with the term "redundancy" in linguistics is deepened by Chomsky's multiple reference to the term referring to the principles of grammar design whether economical or redundant in the formulation of rules (Chomsky 2005: 10). Chomsky's use of the term can be placed at a meta-linguistic level of language description, while the background of Shannon's concept is information transmission seen at a performance level (in

notable attempts at covering all the major features of redundancy are that of Šabršula (1975: 101) and of Slama-Cazacu (1962: 19), underlying specific expressions of redundancy such as the repetition of morphosyntactic information (e.g., noun-verb agreement), or the idea of an excess in signaling elements over the "*strictement* nécessaire". The most complex factor in defining redundancy seems the necessity of integrating the traditional idea expressed in the rhetorical sense of the word with the deeper but linguistically oversimplified view of the concept developed within Shannon's theory.

The definition used in the present work is that proposed in Chiari (2002: 150-151); a text is *redundant* if, on at least one level of analysis, it exhibits properties which diverge from a random text composed of elements functionally distinct, equiprobable, independent from each other, and in which each combination of the elements is a "legal" sequence of the code. Thus we observe redundancy where: a) more than one element from the same level plays the same distinctive role; b) the elements from one level show different frequencies of occurrence, or there are constraints to the combinations of elements in sequences. The former can be called *vertical redundancy*, the latter *horizontal redundancy.*

This definition is an attempt at taking into consideration both the functional aspects of redundancy originated by repetition and the statistical aspects that connect redundancy to predictability factors. Moreover redundancy can be detected both in features of language which are "built-in" in the grammatical structure of the language (*systemic redundancy*) and in features which are freely selected by the user of the language for his concrete speech act (*enunciative*

chomskyan terminology) dealing with the finite nature of the participants in the communication process (speaker and listener) and the equally finite and physical nature of the channel and transmission itself.

redundancy). Systemic or grammatical properties of a language impose a certain amount of redundancy to text in a non avoidable way (non avoidable for that specific language, e.g., grammatical agreement, government, phonotactic constraints, etc.), enunciative characteristics are, on the opposite, chosen by the language user for specific rhetorical purposes, depending on the situational context, as a determination of language use (e.g., double object reduplication in Italian, pleonasm, etc.). Nevertheless the notions of systemic and enunciative redundancy are to be seen as developing in a *continuum*, especially if we observe languages over time.[6]

What are the key consequences and functions performed by redundancy? What are its major manifestations in different

[6] A similar distinction was proposed by Gillette &Wit (1999: 4): "Grammatical redundancy is internal to the language system, is systematic and obligatory, whereas contextual redundancy is voluntary. […] Grammatical redundancy is the internal systematicity and rule governed behavior of a language in which two or more of its features serve the same function." And further on: "Contextual redundancy is the repetition of information that is, in a grammatical sense, non obligatory" (1999: 9). While on the surface the distinction may appear identical, looking more closely there are some major differences. First of all in Gillette and Wit the distinction is categorical, while here we present a more dynamic continuum (which becomes obvious if we observe languages diachronically) for the two typologies. Secondly, Gillette and Wit tend to identify redundancy with repetition only (functional in grammatical redundancy, informational in contextual redundancy), disregarding the role of frequency and probability and the role of transitional properties derived from distributional constraints, which are definitional properties in the characterization presented in this paper. Thirdly, while Gillette and Wit identify the dichotomy with that of obligatory/voluntary, this distinction does not completely coincide with the point of view here assumed, that stressed the "origin" (systemic or enunciative) of the textual properties deriving from redundancy. Besides we cannot properly use the term "voluntary" in most of enunciative cases, we could rather say optional. Among the optional features we can include "voluntary" choices which are determined by rhetorical patterns and preferences. Furthermore, neither the distinction between obligatory and voluntary features is categorical, but defines deeply the dynamic character of natural languages (hints in this direction are presented, for example, in Martinet 1985).

languages?

Briefly, redundancy can play many different roles, all tending towards a better compromise between the speaker's and the listener's needs in a specific context. One of the major functions of redundancy is allowing a sort of storehouse of possible new lexemes, by the non-saturation of all potential (statistically "legal") combinations of phonemes to form words (thus *strufia* is a virtual Italian word, as *tentle* is a virtual English word deriving from non-saturated combinations). Redundancy can also facilitate listening, understanding and speakers' synchronization, through a progressive check of what has been produced and received. It can act as a security system for hypoarticulated speech, through the predictability it confers to the elements of the message and as an error-correcting device. Redundancy generates complexity in linguistic systems, by the imposition of constraints and asymmetry to the functional structures. And finally, it confers compensation (flexibility) to synchronic and diachronic aspects of the language, maintaining or shifting redundancies from level to level.[7]

[7] Dynamism that occurs in the compensation of different redundancy levels is suggested, for example, by Dressler (1969a: 77-78): "Wenn die Redundanz ein essentieller Bestandteil jeder Sprache ist, so folgt daraus deduktiv, dass die Redundanz in der Entwiklung einer Sprache im wesentlichen bewahrt bleiben muss". Pulgram (1983) and Cohen (1969) both observe different aspects of diachronic shifts in redundancy loads from level to level. Cohen notes, for example, how the progressive fall of flexional endings in first, second and third person singular, and third plural of French (*aim*) from the full system of Latin, required an adjustment making the personal pronoun obligatory (as if part of the verb itself) as in *jaim, tu aim (taim), ilaim, élaim, ilzaim, èlzaim* (Cohen 1963: 62-63), that can be interpreted in terms of redundancy balance. A balance between grammatical or systemic factors and enunciative redundancies can be seen as the ordinary process of proportioning redundancy depending on contextual characteristics and pragmatic aspects. This seems a possible interpretation in terms of redundancy of the H&H theory on the relationship of information that is internal or external to the phonetic signal, thus making automatic choices dependent on the richness of the context (Lindblom 1990). A similar position is in

If we accept the definition here proposed we can observe manifestations of redundancy in all natural languages, at different levels, from phonology to morphology, syntax, etc. Constraints in the sequence of elements determine *distributional redundancy*, which can manifest itself in phonotactic restrictions, morphological processes, syntactic rules. For instance distributional redundancy manifested by phonotactic constraints is due to the conditional probabilities governing phoneme sequences (wether we view them as rule-governed or driven by use, thus representing statistical tendencies). This kind of redundancy is the patterning of clustering of phonemes within a language, mainly at the syllabic level. This form of redundancy has also been noted by Brosnahan and Malmberg (1970: 177) at the phonetic level as a possible explanation of how speakers and listeners can communicate dispite ordinary elisions (hypoarticulations), by using the internalized knowledge of those patterns governing co-articulation and sound sequences in each natural language.

An example in Italian is the behaviour of the voiceless plosive [p] that can occur (in hyperarticulated speech): a) at the beginning of a word preceding a vowel #_V, [p]ane (“bread”); b) word initially preceding a consonant #_C [p]rima (“before”); c) between two vowels V_V, a[p]e (“bee”). It cannot (or simply does not) occur a) at the end of a word, except for loan words, _#, *...p; b) after a plosive b_, *bp; c) between two plosives C_C, *tpg, *kpt (Scalise, 1994: 31). The non casual distribution of phonemes in a language generates the power of predicting sequences, and reconstructing sequences in the event of noise.

Another example is the redundancy of consonant clusters in

fact taken by Pulgram (1983: 109): “It would be pleasant to report that either preservation or elimination of redundancy, that is, its longevity, is proportionate to its importance for the intelligibility of an utterance”.

Italian (Chiari 2002: 213-233; Chiari & Castagna 2004). If we take triconsonantal clusters, at the abstract level we can permute the 19 consonants of Italian to form 5,814 clusters, being in fact 11,828 if we double the positions considerinf separately initial and middle position. In a corpus of spoken Italian (Lessico di frequenza dell'italiano parlato, section recorded in Rome) consisting of 129,056 words, we observe only 40 triconsonantal clusters, consequently having a redundancy of:

(3) 1 - (40/11,828) = 0.9966

Another redundancy type derives from the presence of elements playing the same functional role. S*yntagmatic redundancy* is represented by phenomena like government and *functional redundancy* is represented by agreement, pronoun reduplication and double negation.

Syntagmatic redundancy originated from sequential constraint can be found at the morphological level (e.g., in restrictions governing the application of affixes to a base) and at the syntactic level: in word order, but also as Pierre Guiraud (1968: 164) noted in the mere expectation of a noun following the sentence:

(4) Yesterday I went to play ….

Syntax itself in Guiraud's words does not exist without redundancy. The most evident manifestation of this sort of redundancy is government, where an element takes a determined category from another element in the syntactic construction (even though, unlike in agreement, categories are different are presented by the two elements). Government provide predictability to the sequences as we can see, for example in case government in latin or russian:

(5) Слушать радиопередачу (*clušat' radioperedaču)*
'listen to a radio programme'

In the previous example the verb *clušat* "to listen" requires a pronoun or noun in the accusative case, thus imposing the category to the following element.

(6) Перевёл книгу с русского языка на английский
Perevël knigu s russkogo jazyka na anglijskij
'translated the book from Russian to English'

In the example in (6) the verb *perevodit'* 'translate' governs the accusative case of *kniga* 'book'*,* the phrase starting with preposition *s* + genitive of *jazyk* 'language' (which concords with adjective *russkij* 'Russian'), and the phrase starting with preposition *na* governing the accusative of *anglijskij* 'English'. Redundancy of government is dependent on the number of controlling elements (C_1, C_2, …C_n), and the number of depending elements (D_1, D_2, …D_n) , present in a corpus of texts. In ten samples of the same length extracted from the Russian translation of *The Name of the Rose* by Umberto Eco[8] this kind of redundancy is:

(7) $$R_{russ} = \frac{\sum D_1 D_2 ... D_n}{\sum C_1 C_2 ... C_n} = 1.48$$

The interpretation that follows is that, since in a language that uses government the value cannot be lower than 1, in the text chosen

[8] The choice of *The Name of the Rose* by Umberto Eco was made to be able to confront the same text in different languages, especially for the redundancy of agreement analysis based on Italian, French, and Russian reported in more detail in Chiari (2002: 271-294).

there is generally more than one depending element determined by one controller.

As an example of *functional redundancty* we can analogously observe agreement which operates in a similar way to government copying the same feature posessed by a controller to a depending element (Corbett 1991). Romance languages (8) for example exhibit gender and number agreement for nouns (Repina, 1996: 18), while Russian (9) shows gender, number and case (Chiari, 2002: 281):

(8) *Fr*. *Un* veu*f* malhere*ux* *m*. / *Une* veu*ve* malheur*euse* *f*.
It. *Un* pover*o* vedov*o* *m*. / *Una* pover*a* vedov*a* *f*.
Sp. *Un* viud*o* desconsolad*o* *m*. / *Una* viud*a* desconsolad*a* *f*.
Port. *Um* viúv*o* velh*o* *m*. / *Uma* viúv*a* velh*a* *f*.
Rum. *Un* vădu*v* bătrî*n* *m*. / *O* văduv*ă* bătrîn*ă* *f*.
‘A sad widower / A sad widow’

(9) Rus. Длинн*ая* улиц*а* (*dlinnaja ulica*) *f*.
‘long street’
Зелён*ое* дерев*о* (*zelënoe derevo*) *n*.
‘green tree’
Стар*ый* журна*л* (*staryj žurnal*) *m*.
‘old magazine’

A further typology of redundancy is *statistical redundancy*, occurring at phonological, lexical, syntactic levels, deriving from constraints in the language code that determine a specific statistical profile to the texts produced in each particular language and that let the listener predict different portions of the spoken (or written) chain by accessing their competence about language patterns.

3. Redundancy in Artificial Languages

Many have argued that redundancy should be eliminated from artificial languages, aiming at a better economy of means. But can it be eliminated? To what extent? What kinds of redundancies are there? Which can be eliminated? Is there a price for this loss?

Different opinions, not only on the effective role of redundancy in historical linguistics, but also on the interpretation of the role it plays in languages, and specifically in artificial languages is touched upon briefly by Dressler (1969b) and Pulgram (1983).

Pulgram observes how useful it is to investigate how artificial and international languages deal with the problem of redundancy. Grammatical redundancies, their presence and degree, for Pulgram, have generally been neglected in Esperanto, Ido, Volapük and Basic English, from the theoretical and practical point of view. Moreover Pulgram envisages redundancy in artificial languages as a negative property: "one is astonished that many of these languages, basically Indo-European and mostly Latin-Romance in structure, so often insist on totally useless redundancies – e.g., gender, an intricate system of grammatical but non-signaling concord" (Pulgram (1983: 110). The position assumed by Pulgram appears of great critical awareness, mainly on some problematic aspects about the evaluation of how to determine when redundancy plays a role in linguistic change and in facilitating actual communication contexts, although it does not accurately face the problems concerning auxiliary, artificial and planned languages.

On the contrary Dressler (1969b), in a brief essay completely devoted to redundancy in artificial languages, *Zur plansprachlichen Redundanz,* underlies the fact that since redundancy plays a capital role in natural languages, it should also be a central topic of discussion in planning artificial languages. Dressler claims that a certain amount of syntagmatic redundancy deriving from the different frequency exhibited by elements of the same level of analysis in combination can be appropriately excluded. This kind of redundancy is strictly dependent on the non-saturation of possible combination of elements (what we have called *distributional redundancy*) and is deeply related to the structure of the linguistic system itself.

Aiming at maximum economy, the language planner can avoid some syntagmatic redundancies to achieve a more flexible usage, but it generally does not avoid *paradigmatic redundancy*, which depend on the inventory of the system, the combination of the elements and their degree of actual usage. Dressler stresses the presence of elements of redundancy that are hard if not impossible to expunge from the system, since they are closely connected to the usage the speech community makes of the language. This position echoes Saussure's observation (1916: 111) on the inevitability of change in artificial languages.

Artificial languages, if spoken, are subject, as any natural language, to noise factors that are intrinsic to communication activities, and thus cannot be removed without some loss in the outcome of the performance. A language without those aspects of redundancy would be useless from a practical point of view. Relevant to this argument appear the words of Guiraud (1969: 165) on international languages, being by necessity characterized by an high coefficient of redundancy when spoken as a means of avoiding noise. This implies the paradox that the higher the level of redundancy the faster the process of change and variation will occur.

This standpoint tends to contradict by means of the argument of communication needs and noise-reduction functions of redundancy the possibility of eliminating redundancy from artificial languages. The claim is also related to the problem of placing the role of redundancy in the evolution of languages. In particular if languages possess special varieties of redundancy at all planes (as Dressler, 1969a, Pulgram 1983, Cohen 1963; 1969 show with exemplifications) and if the level of redundancy of one plane is related to the other with some causal diachronic link (even if this causality can be hard to sustain), it is not unlikely that the contingently reduced redundancies could reappear in other forms on a different plane.

4. Some Examples

4.1 Redundancy of Graphemic Systems

As an example of statistical redundancy, we will observe and discuss the values of redundancy of the graphemic system of an artificial language, compared to that of a number of natural languages. Dressler (1969b) offers values for zero and first order entropy for Esperanto, following the formula proposed by Shannon (Shannon & Weaver 1949: appendix 2):

$$(10)\quad \boldsymbol{H = -k\sum_{i=a}^{z} p_i \log p_i}$$

Where *k* is a constant, depending on the choice of measurement unit. The formula states that information (H, entropy) is equal to minus the sum of the probabilities (p_i) of each of the elements times the logarithm of p_i. Redundancy depending on relative entropy will be measured by:

$$(11)\quad \boldsymbol{\rho = 1 - \frac{H}{H_{\max}}}$$

Where maximum entropy (H_{max}) corresponds to the case of all equiprobable symbols. Thus redundancy is one minus relative entropy (Shannon & Weaver 1949: 45). From 1,000 grapheme samples, with a set of 25 graphemes, Dressler (1969b: 276) obtains for Esperanto:

(12) $\boldsymbol{H_{ESP} = 4.0893}$

And a relative entropy of:

(13) $\boldsymbol{h_{ESP} = 0.8849}$

Thus redundancy of the first order grapheme system of Esperanto is:

(14) $\boldsymbol{\rho_{ESP} = 1 - h = 0.1151}$

The result achieved indicates a slightly lower value to that of German, as computed by Meyer-Eppler (1952), of $\rho_{GER} = 0.15$ bits per symbol. From the graphemic point of view Dressler concludes that the level of redundancy of Esperanto is adequate to what expected for any *a-posteriori language*, that is a language having as a source (mainly for vocabulary) existing natural languages.

Nevertheless, besides the facts that values obtained depend on specific linguistic choices made (e.g. the number of graphemes depends on how diphthongs have been treated, how spaces and punctuation have been accounted for), if we compare these results to that achieved by Manfrino (1960), Moles (1958), Küpfmüller (1954), Barnard (1955), Shannon (1951) for other languages, using the same method of computation, the comparison does not appear so significant.

Table 1. Information and Redundancy of First Order Approximation

	H_0	H_1	h	ρ
Italian (Manfrino 1960: 11)	4.39	3.95	0.90	0.10
Hebrew (Moles 1958)	4.46	3.99	0.89	0.10
Esperanto (Dressler 1969b: 276)	*4.64*	*4.09*	*0.89*	*0.12*
English (Manfrino 1960: 26)	4.70	4.13	0.87	0.13
German (Manfrino 1960: 26)	4.70	4.08	0.87	0.13
German (Küpfmüller 1954:70; Barnard 1955: 51)	4.76	4.10	0.86	0.14
English (Shannon 1951)	4.76	4.03	0.85	0.15
French (Manfrino 1960: 26; Barnard 1955: 51)	4.70	3.98	0.85	0.15
Spanish (Manfrino 1960: 26; Barnard 1955: 51)	4.70	4.00	0.85	0.15

Not only if we observe data in Table 1 we discover that Italian and Hebrew possess an inferior degree of redundancy in their graphemic systems to that of Esperanto, but we also note that values of the immediately following languages are placed in a continuum.

On the other hand data presented refer only to first order approximation to the languages analyzed, that is to say that only relative frequencies of each graphemes are taken into consideration, without any account of transitional probabilities involving sequences of graphemes (that might roughly represent phonotactical constraints). Thus data is incomplete and cannot actually give an accurate picture of statistical redundancy, nor of its relation to distributional redundancy.

4.2. Phonotactic Patterns and Frequency Effects

As we have seen in the previous paragraphs, distributional redundancy derives from constraints in the sequence of elements presented in the messages produced by a specific code. At the phonological level rules or tendencies that govern restrictions in syllable formation, tendencies for the onset or rhyme, produce syllabic distributional redundancy (De Mauro 2003: 35-36, Chiari 2002: 240-245). Phonology rarely has been at the centre of language planning reflection, although the tight relationship between most planned languages to the Indo-European family determine evident similarities in the preference for CV structures, and in the choice of some specific restrictions (Oosterdorp 1999: 52). Esperanto, for example, exhibits a syllable structure extremely close to Italian, which has a high syllabic redundancy (both at vocabulary and textual levels). In Italian the five most common syllable structures (CV, CVC, V, VC, CCV) cover about 97% of syllable occurrences in the vocabulary, and between 90.7 and 98.6% of occurrences in texts. Redundancy varies from 0.58-0.62 for the most common syllables, to 0.42-0.50 considering all attested syllables (including loanword structures) (Chiari 2002: 244-245).

Sillable structure in Esperanto is very similar to that of Italian. If we observe, for example, consonant clusters occurring as onsets in esperanto syllables, the constraints governing phoneme sequences are analogous to Italian two consonant clusters: “The first segment always has to be an element of the set {b, d, f, g, k, p, s, š, t, v} and the second one an element of {r, l, n}” (Oostendorp 1999: 57). The similarity applies also to three consonant onsets starting with *s*, while in Italian the fricative represented by *š* cannot be placed in a three consonant onset.

It has been often pointed out that, in terms of simplicity of articulation, consonant clusters constitute an obstacle for some speakers, thus leading to the idea of eliminating or limiting the use of clusters from planned languages (this is the perspective taken by R. Harrison in his guidelines for the design of an optimal international auxiliary language). The notion of articulatory complexity is a particularly problematic one, since parameters of complexity have not been properly detected.[9] Common proposals tend to coincide with the choice of the first syllable structures produced by infants of all languages. A reflection that takes into account common phonetic contexts that produce variation (and change over time) is still missing (e.g., not only partial assimilation place of articulation, of mode and of sonority, co-articulation – systematic and non systematic phenomena - in particular that regarding sequences of vowels, etc.). Issues regarding articulatory complexity should certainly be part of planned AL design and would shape distributional properties of syllables. Further reduction of consonant clusters will evidently directly affect distributional redundancy, by making sequences more predictable.

4.3. Allomorphy and Suppletion Reduction

Among features of planned languages a common issue is the tendency towards the elimination of allomorphs (that might be present in source languages), excluding or leveling by analogy cases like the English plural *–s, -es, -en* or vocalic alternation (e.g. *-s* in *lists*, *-es* in *houses*, *-en* in *oxen*, *men*, *geese*), and the extreme case of suppletion (where a root is substituted with another, like in *good/better*, or in Italian *andare/vado*).[10] This is a planning rule

[9] André Martinet was one of the first to propose the idea that complexity of articulation could be subject to principles of economy, following the indications present in Zipf (1935: 1949).

[10] "For the signs X and Y to be suppletive their semantic correlation should be maximally regular, while their formal correlation is maximally irregular." (Mel'čuk 1994: 358).

which aims at fighting redundancies deriving from the asymmetry between function (e.g., morpheme for the plural) and forms (e.g., the signifiers that characterize the different morphs realizing the morpheme), which we called *functional redundancies* (*vertical*).

Esperanto explicitly aims at eliminating allomorphy from roots (Wells 1978). Interlingua on the contrary presents multiple cases of allomorphy regarding roots. Bernasconi (1977) observes how this difference is manifested throughout the two languages.

Table 2. Roots for the Action of Speaking (Gilbert 1962: 31)

Esperanto	Interlingua	Family of	
parol-o	parol-a	*parol*:	Paroletta (wordlet) parola de honor (word of honor)
parol-a	verb-al	*verb*:	verbalisar (to verbalize)
parol-e	or-al-mente	*or*:	oralitate (mouth-ness), ore (mouth) orificio (orifice)
parol-i	parl-ar	*parl*:	parlamento (parliament) parlatorio (place for speaking) altoparlator (loudspeaker)

If we give a look at the respective vocabularies for Esperanto and Interlingua the picture is confirmed. Though cases of suppletion are more common in Interlingua there are some examples in Esperanto too. When we give a look at confixes such as those used to express concepts related to water, we find that both Esperanto and Interlingua also tend to use the Greek forms exspecially in technical terms. Esperanto uses *akvo, akvaforto, akvario, akvorezista, akvoĵeto, akvomasaĝo, akvomotoro* and also *hidranto, hidraŭliko, hidrocefalo, hidrofobio, hidrologia*. Interlingua uses *aqua, aquar, aquarella, melon de aqua* and also *hydrante, hydraulic, hydric, hydrocephalo.* While Interlingua has a suppletive triplet *cavallo /equin / hippic,* Esperanto keeps *ĉevalo / ĉevala.* In some cases in the vocabulary slips the Latin influence, as for the “mother” lexical items, *patrino, patrineco, patrinece* which coexist with *matriarka, matriarkeco.*

The different choice here lies in the typological structures of the languages analyzed. Esperanto has a strong tendency towards highly regular agglutination, while Interlingua, being robustly connected to lively neo-romance languages exhibits strong flexional morphology, thus presenting a large number of inherited allomorphies.

Is this sort of redundancy avoidable? Answers are manifold. Yes, if Esperanto succeeds in manifesting a tendency at reducing allomorphy it must be possible, at least in the design, to expunge this feature. The difference in allowing or prohibiting allomorphy is (and should be) served to the main purpose of the planned language, which in the case of Esperanto and Interlingua is obviously different. In natural languages suppletion is connected with many factors: it tends to be preserved in frequent words, it tends to appear more frequently in inflecting languages than in agglutinative languages. Since Esperanto is tendentially agglutinative it is expected to present a smaller number of cases of suppletion.

Furthermore allomorphy and suppletion are represented by a wide variety of cases some of which of theoretically difficult analysis, exspecially when we deal with abstract words.

The question about the elimination of allomorphy is furthermore connected to the way one envisages the role of linguistic change that touches (or might touch) artificial languages (§3). Observing natural languages diachronically the emergence of different forms to represent the same root determining morphological alternation (like metaphony in German) is a frequent consequence of conditioned phonetic change (or a sequence of phonetic changes that leave the original context of change non-transparent). If transmission of a language happens mainly by written communication the process of change is slower, than when the language is spoken by a large community. Under this light certainly artificial languages tend to be more conservative. This conservativeness does not depend entirely on the purpose and the plan of the language (as it would be desirable in the planning perspective), but also on the natural forces that contribute to linguistic change in ordinary contexts. Even though change is slower in artificial languages, it does indeed occur (as has been shown for Esperanto by Herring 2005; Nicholas 2002), as well as variation occurs (Sherwood 1982b).

4.4. Grammatical Agreement and Government: Functional Redundancy

At the morphological and syntactic levels a condition of asymmetry in grammatical functions and forms produces what we have called *functional redundancy*. Examples of this typology are agreement or concord (Italian gender-number for nouns, determiners and adjectives, like in *la bella ragazza*, object reduplication like in *il caffè lo prendi?*, double negation, like in French *Ce n'est pas rien*) and government.

Agreement is almost unanimously considered an expression of redundancy where the same feature (mostly grammatical) is repeated (copied) to different elements syntactically linked. Features that express agreement have also been interpreted as discontinuous signifiers, bearing the same function (Martinet 1985: 60-61). The redundant character of this phenomenon can be easily exemplified by the difference in agreement signaling which distinguishes French oral and written language: spoken language being generally less redundant (*je aime, tu aimes, il aime, ils aiment* being pronounced homophonous) than its corresponding written form. Šabrušula (1975) notes how the lack of functional marks in oral speech does not affect the success of communication, since redundancies at other levels compensate for the elimination of information at morphological level.

In a similar way, most languages with a rich flexional system (like Italian, Spanish, Russian, etc.) let hypoarticulated speech delete or neutralize phonetic distinctions even when they affect relevant morphological markers, that become indistinct. The predictability of agreement features in a language does in fact allow speakers to reconstruct (if necessary) deleted portions of the speech chain.

For a set of samples from *The Name of the Rose* by Umberto Eco both in its original Italian and in its French translations measuring redundancy from gender and number agreement following (7) we obtain (Chiari 2002: 285):

Table 3. Gender and Number Redundancy in Italian and French (*The Name of the Rose* Samples)

	Italian	French
Gender redundancy	1.46	1.33
Number redundancy	1.86	1.55

Planners and theoreticians of artificial languages have often indicated agreement (e.g. gender and case, but also person and number in finite forms of verbs, above all) as a feature that should be avoided. In effect Sherwood (1982a: 5) reports that "the Esperanto accusative has often been attacked as being excess baggage and inappropriate in a language intended to be easy to learn and use [….] Perhaps the most sophisticated defense is the one which points out that in communication between speakers from different cultural backgrounds, extra precision and redundancy are needed to compensate for the lack of shared assumptions and backgrounds".

Esperanto, for example, avoids the common romance agreement of determiner (article) and noun, through a sole definite form (*la*), independent from gender and number of the noun, preventing texts from displaying duplicated markers in noun phrases. On the contrary Esperanto preserves noun-adjective agreement in case and number (which Ido eliminated), even though the question about the opportunity of preserving case is still a matter of debate:

(15) la dolĉa gitaro
‘the sweet guitar’ N. sing.
(16) La dolĉaj gitaroj
‘the sweet guitars’ N. pl.
(17) Mi ludas mian dolĉan gitaron
‘I play my sweet guitar’ Acc. sing.
(18) ili ludas iliajn dolĉajn gitarojn
‘They play their sweet guitars’

In the examples above it is possibile to observe the repetition of the number and case information of the noun and in the adjectives and possessives.

As an example of redundancy of agreement in esperanto a sample from the speech of Zamenhof at Boulogne sur mer (1905) was chosen. The sample consisted of 229 words of the speech. If we consider the controllers (C=33) for number (nouns) and the determined elements (D=47), the agreement redundancy of the text results in:

$$(19)\quad R_{esp} = \frac{\sum D_1 D_2 \ldots D_n}{\sum C_1 C_2 \ldots C_n} = 1.42$$

Since values are determined not only by the features of agreement as a linguistic property of the language, but are also dependent on the specific text under analysis, and on its peculiar style it would be interesting to compare parallel texts of different typologies.

To be able to evaluate cross-linguistic differences we chose to as a sample a paragraph (111 words) from the beginning of the novel *Bouvard et Pécuchet* by Gustave Flaubert, in its original French and in Italian, English and Esperanto translations), here is what we obtain:

Table 4. Number Agreement in a Sample from *Bouvard et Pécuchet*

	French	*Italian*	*English*	*Esperanto*
Nr. Words	111	111	118	110
Controllers	*23*	*23*	*3*	*10*
Determined	39	39	3	11
Number Redundancy	*1.69*	*1.69*	*1*	*1.1*

The second sample is taken from the biginning of the first chapter of *Alice in Wonderland* by Lewis Carroll (253 words), in the original English and in French, Italian and Esperanto translations.

Table 5. Number Agreement in a Sample from *Alice in Wonderland*

	English	*Italian*	*French*	*Esperanto*
Nr. Words	253	207	292	203
Controllers	*6*	*32*	*52*	*12*
Determined	6	51	69	13
Number Redundancy	*1*	*1.59*	*1.33*	*1.08*

The picture showed by the comparison clearly meets the expectations. While (written) French and Italian are the most redundant in number category signalization (because it involves adjectives, articles and verbs), English is at the extreme opposite (since it does not posses any agreement for number in noun-adjective relationships, nor a difference in the definite article, and for verbs the difference emerges only for example in third person singular of present tense). Esperanto lies in the middle of the two extremes: the value of redundancy is low, because for each controller there tends to be only one determined element, but the general number of controllers is higher than that of English since Esperanto allows noun/adjective agreement. Another variable is text length. As can be seen in Table 5 translations of the same text, expecially for literary texts can vary significantly introducing elements that determine a change in the representation of number redundancy. The situation presented by the French translation is evident. The text is definitively longer than the original English (while Italian and Esperanto are shorter) and presents a greater amount of controllers if compared to Italian, partly because has the obligatory pronoun subject expression which is not a feature of Italian. Still Italian has the highest redundancy having a greater mean number of determined elements per controller.

Oostendorp (1998) reports another typology of redundancy appearing in suffixes such as in the case of the word *samseksemulo* both the suffixes *–ul* and *–o* indicate a nominal category. Oostendorp attributes to the violated principle of representational economy the necessity of introducing abbreviated forms that appear non redundant in colloquial Esperanto used in informal speech: "shortening is only triggered by a form of morphological economy of representation: the same categorial feature cannot be expressed more than once within the same word" (Oostendorps, 1998: 185). Moreover the elimination of some forms of redundancy seems to be connected with spontaneous Esperanto speech, more than with its written form.

Another aspect of functional redundancy is the phenomenon of government, or rection, where "the difference between concord and government lies in the fact that under concord two or more words or phrases are 'inflected' for the same category (e.g., number or person), whereas under government the *principal* and the *dependent* member of a syntactic construction do not both exhibit the same category: instead the dependent member is determined with respect to the relevant category (e.g., case) by the principal member" (Lyons 1968: 241). Redundancy here comes from the predictability of the category imposed on the dependent member. Artificial languages tend to avoid this functional redundancy by stating as a rule the absence e.g. of prepositions determining case on nouns.

Esperanto contemplates only one case of government, that of preposition indicating movement:

(20) Mi iras en la urbon
'I go to town'

This kind of government is a weak form since the same preposition can actually govern nominative in sentences not expressing motion as in :

(21) mi loĝas en Italujo
'I live in Italy'

The predictability is thus produced by the combination of the verb (and its semantic features) and the preposition, determinig the case to be applied to the nouns and adjectives following.

4.5. Word Order Redundancy and Statistical Properties

At syntactical level, artificial languages have often dealt with word order issues. Word order manifests different degrees of distributional redundancies depending on restrictions imposed to sequences. Against free word order in artificial languages are those who suggest that free order would necessarily imply a case system, that is generally perceived as a complex learning feature of languages. The presence of restrictions (often associated with absence of cases) determines distributional redundancy at syntactic level, by limiting the number (an frequency) of possible combination of the elements of the inventory. [11]

The SVO order (with head preceding modifiers) is generally

[11] The association of the loss of Latin case system in Romance languages like French, Spanish or Italian to the emergence of constraints in word order is quite common. Pulgram (1983: 114-116) interprets this linguistic change as a result of a process of displacement of redundancies from a morphosyntactic level to word order syntax.

considered the most advisable in artificial languages, being "natural", and most used in the languages of the world. This choice instead of removing syntactic redundancy increases it (as in phonotactic distributional redundancy), and determines a wider possibility of predicting sequences.

Herring (2005) observes how: "In theory, Esperanto sentences can be scrambled in any possible way. In reality, however, there are any number of minor rules which specify which items should precede or follow given other types of items. Prepositions and determiners, for example, *always* come before words they specify, and the same is true in the majority of cases for adverbs as well". The introduction of rules for sequences determines the emergence of an extremely common manifestation of redundancy, which is coupled to the more general statistical properties of lexical sequences characterizing a language. Surprisingly, Esperanto shows peculiar statistical properties which, though similar to those of Italian, Spanish, French and German (more distant from English), make it easily detectable by machine learning techniques, as Manaris et al. (2006) demonstrate. The authors claim that: "Esperanto, in spite of its short, 120-year lifecycle, has evolved enough to exhibit "natural" statistical proportions […], analysis of misclassified patterns indicates that Esperanto's statistical proportions resemble mostly those of German (59.1%) and Spanish (27.3%), followed by English (11.4%) and Italian (2.3%)" (Manaris et al. 2006: 107).

It is interesting here to note that the similarity from the statistical point of view surprisingly indicates a closer likeness of Esperanto to German than to Italian (which, on the contrary is qualitatively closer to Esperanto even for its syllabic distributional features). This indicates autonomy of different levels in terms of redundancy measure.

5. Conclusion

As Sapir (1925) pointed out an artificial language, besides being simple, regular, and logical, needs to be rich and creative. But richness and creativeness are based heavily on some principles of redundancy (at systemic and enunciative levels). In this paper we have concentrated our attention on systemic expressions of redundancy, although a long series of phenomena occur at the enunciative level.

Some examples of how artificial languages have dealt with redundancies have been presented. When redundancy is perceived as a negative property, solutions have been proposed. Esperanto, Interlingua, Ido present all a number of systemic redundancies, some of which are not avoidable, mainly because they constitute the structure of the language itself. Others, like functional redundancy in agreement and government, have been object of a specific reflection and have been reduced or eliminated to different extents. A direct consequence of this reduction is a more rigid word order, increasing redundancy at the syntactic level.

Even though the claim that redundancy is globally preserved by languages, like Dressler and Pulgram suggest, is empirically not verifiable or falsifiable, it is an extremely appealing interpretation of linguistic change in a non-teleological perspective. Even artificial languages do not escape this picture. No artificial language has expunged redundancy of all kinds. A language completely deprived of redundancy would not even have the form, needless to say the usage and communicative functions, of a language.

It seems as if a language in use (whether planned or natural) needs a certain amount of tolerance of redundancy, if it is to be expressed (spoken, written or in any other form) for communicative purposes by living creatures (humans and non humans). Redundancy is thus a necessary tool for psycho-biologically finite users.

References

Barnard, A. 1955. Statistical Calculation of Word Entropies for Four Western Languages. *IRE Transactions in Information Theory* 1, 49-53.

Bernasconi, E. 1977. The Neo-Romance Languages. *Esperanto aý Interlingua* 86-110**.** La Chaux-de-fonds: Kultura Centro Esperantista.

Chiari, I. 2002. *Ridondanza Linguaggio. Un Principio Costitutivo delle Lingue*, Rome: Carocci.

Cohen, M. 1963. Problème de la surabondance dans le langage. *Le Courrier Rationaliste* 4, 60-66.

_____. 1969. Quelques vues sur les équilibres linguistiques. In J. Dierick & Y. Lebrun (eds.), *Linguistique contemporaine: Hommage à eric buyssens*, Éditions de l'institut de sociologie, Paris, 19-25.

De Mauro, T. 1998. *Linguistica Elementare*. Bari-Roma: Laterza.

Dressler, W. 1969a, Die Erhaltung der Redundanz. Lateinische Beispiele für ein Wenig Beachtetes Prinzip der Sprachentwicklung. *Studia Classica et Orientalia Antonino Pagliaro Oblata* 2, 73-84.

_____. 1969b. Zur Plansprachlichen Redundanz. *Österreichischen Akademie der Wissenschaften* 106, 274-80.

Ellis, N. 2002a. Frequency Effects in Language Processing. A Review with Implications for Theories of Implicit and Explicit Language Acquisition. *Studies in Second Language Acquisition* 24, 143-188.

_____.2002b. Reflections on Frequency Effects in Language Processing. *Studies in Second Language Acquisition* 24, 297-339.

Gilbert, W. 1962. *Planlingvaj Problemoj*. La Laguna: Régulo.

Gillette, M. & Wit C. 1999. What is Linguistic Redundancy. *Technical Report*. Chicago, IL: University of Chicago.

Grice, P. 1975. Logic and Conversation. In P. Cole & J. Morgan (eds.), *Syntax and Semantics* 3, 41-58. New York: Academic Press.

Guiraud, P. 1968, Langage et Théorie de la Communication. In A. Martinet (éd.), *Le Langage* 145-68. Paris: Encyclopédie de la Pléiade.

Herring, J. 2005. Syntactic and Lexical Changes in Esperanto: A Corpus-based Survey. *2nd Midwest Computational Linguistics Colloquium*.

Küpfmüller, K. 1954. Die Entopie der Deutschen Sprache. *Fernmeldetechnische Zeitung* 7, 265-272.

Lindblom, B. 1990. Explaining Phonetic Variation: A Sketch of the H & H Theory. In W. Hardcastle & A. Marcha (eds.), *Speech Production and Speech Modelling* 403-439. Dordrecht: Kluver

Academic Press.

Lyons J. 1968. *Introduction to Theoretical Linguistics*, Cambridge: Cambridge University Press.

Manaris, B., L. Pellicoro, G. Pothering & H. Hodges. 2006. Investigating Esperanto's Statistical Proportions Relative to Other Languages Using Neural Networks and Zipf's Law. *Proceedings of the 24th IASTED International Conference on Artificial Intelligence and Applications*, 102-108.

Manfrino L. 1960. The Entropy of the Italian Language and its Computation. *Alta Frequenza* 29, 4-29.

Martinet, A. 1955. *L'Économie des changements linguistiques*, Berne: Franck.

_____. 1985. *Syntaxe Générale*. Paris: Armand Colin.

Mel'čuk, I. 1994. Suppletion: Toward a Logical Analysis of the Concept. *Studies in Language* 18, 339-410.

Meyer-Eppler, W. 1952. Informationstheorie. *Die Naturwissenschaften* 39, 340-52.

Moles, A. 1958. *Théorie de l'information et perception esthétique*. Paris : Flammarion.

Nicholas, N. 2002. Folk Functionalism in Artificial Languages: The Long Distance Reflective Vo'a in Lojban. *Journal of Universal Language* 3, 133-167.

Oostendorp, M. 1999. Syllable Structure in Esperanto as an Instantiation of Universal Phonology. *Esperantologio Esperanto Studies* 1, 52-80.

Pulgram, E. 1983. The Reduction and Elimination of Redundancy. In F. Agard *et al.* (eds.), *Essays in Honor of Charles Hockett*, 107-125.

Šabršula J. 1975. Redondance et Économie. *Acta Universitatis Carolinae, Philologica, Romantica, Pragensia* 9, 101-24.

Sapir, E. 1925. The Function of an International Auxiliary Language. In H. Shenton, E. Sapir & O. Jespersen (eds.), *International Communication: A Symposium on the Language Problem*, 65-94.

Saussure, F. 1916. *Cour de linguistique générale*, Edition critique préparée par tullio de mauro. Paris :Payot.

Shannon C. 1951. The Prediction and Entropy of Printed English. *Bell System Technical Journal* 30, 50-64.

Shannon, E. & W. Weaver. 1949. *The Mathematical Theory of Communication*, Urbana, IL: University of Illinois Press.

Sherwood, A. 1982a. Statistical Analysis of Conversational Esperanto with Discussion on the Accusative. *Studies in the Linguistic Sciences* 12. 1, 165-182.

_____. 1982b. Variation in Esperanto. *Studies in the Linguistic Sciences* 12. 1, 183-196.

Slama-Cazacu, T. 1962. L'économie et la redondance dans la communication. *Cahier de linguistique theorique et appliquée* 1, 17-25.

Wells, J. 1978. *Lingvistikaj Aspektoj de Esperanto*. Rotterdam: UEA.

Zipf. K. 1935. *The Psycho-biology of Language: An Introduction to Dynamic Philology*, Boston, MA: Houghton Mifflin Co.

_____. 1949. *Human Behavior and the Principle of Least Effort*, Boston, MA: Addison-Wesley Press.