Studying restrictions on patterns of word-formation by means of the Internet

Franz Rainer

The Internet has not been conceived as a tool for linguistic research, but being a huge collection of texts it has turned out to be also an important source for linguists. Many studies exploiting this new source of information have been published over the last few years, and many more will undoubtedly see the light in the near future. The present study explores in how far Internet data may be fruitfully used to gain new insights about the nature of restrictions on patterns of word formation. As a test case, I have chosen the Italian intensive suffix *-issimo*, which I had already studied some twenty years ago on the basis of a corpus of a hundred novels and, to a certain extent, through the elicitation of acceptability judgements from native speakers. It is shown that, in the case of a highly productive suffix like *-issimo*, the absence from the Internet may be interpreted as a good indicator of low acceptability, an inference that is generally not allowed by smaller corpora. Another advantage of the Internet data is that it nicely brings to light the gradual nature of restrictions on patterns of word-formation.[1]

1. The Internet as a research-tool for the linguist

As is well-known, the Internet was originally designed as a means of making military communication less vulnerable and then put to use in scientific, commercial and private communication. None of those who developed this new medium ever had the slightest intention of facilitating linguistic research, but since the Internet constitutes a huge collection of texts it didn't take long for linguists to discover its immense potential as a research tool. It is not my intention here to review the numerous ways the Internet has already been put to use in linguistic research over the last few years, but to explore one more possible application which, to the best of my mind, has not yet been exploited.

It will be our goal to explore in how far data gathered from the Internet may be used as a substitute for acceptability judgements in word-formation and whether such data may help to improve the description of restrictions on patterns of word-formation. One of the notorious disadvantages of corpora is the fact that they contain no negative evidence. Now, the Internet is so huge a corpus that one might suspect that the absence of certain instantiations of at least

Rivista di Linguistica, 15.1 (2003), p. 131-139 *(ricevuto nel giugno 2003)*

highly productive patterns should turn out to be a reliable indicator of their low acceptability.

2. The Italian suffix -issimo: the state of the art

The Italian suffix *-issimo* has been chosen as a test case for two reasons: on the one hand, it is a highly productive suffix,[2] and on the other it allows us to compare the results of our study of Internet data with those reached in Rainer (1983a:56-61; 1983b) on the basis of a traditional corpus study complemented with unsystematic elicitation of acceptability judgements. Among the more puzzling results of my previous study were contrasts in acceptability like *contrarissimo* 'fiercely opposed' *vs* ??*letterarissimo* 'highly literary', *tragicissimo* 'extremely tragic' *vs* ??*caratteristicissimo* 'highly characteristic', *attivissimo* 'extremely active' *vs* ??*significativissimo* 'extremely significant', etc. Since no neat correlation could be established with intensifiability, length or suffix/final string of the base, I then concluded that the determining factor for the low acceptability of certain formations was their learnèd character (see Rainer 1983b:101).

From the few more recent studies on *-issimo* I would like to mention only Wierzbicka's (1991:271) observation that this suffix "involves a self-evident exaggeration; and [that] this exaggeration is functional, in view of the speaker's emotional attitude."[3]

3. A successful probe

In order to test the usefulness of the Internet for our purpose, as a first step the frequency of the adjectives in *-issimo* (henceforth *Fi*) quoted above was controlled on the Internet. The results of this probe were encouraging. It turned out that the three acceptable formations were all attested, while none of those with low acceptability was. But since absolute frequencies may be determined by many factors which have nothing to do with the grammar of *-issimo*, I decided to take into consideration also the frequency of the corresponding collocation *molto* 'very' + adjective (henceforth *Fm*) and the quotient of both frequency measures, which we will call the propensity to accept *-issimo* (henceforth *Pi*; Pi = Fi / Fm). By taking into account Fm we make sure that a certain adjective is intensifiable, which is a necessary condition for the use of *-issimo*. And by means of Pi we may measure not only whether a certain adjective qualifies as a base of *-issimo* or

not, but also the degree to which it may be called a typical base of our suffix. In our six test cases, e.g., *contrarissimo* (Pi 1.22 = Fi 28 / Fm 23) and *attivissimo* (Pi 0.77 = Fi 895 / Fm 1,169) turn out to be more typical bases than *tragicissimo* (Pi 0.22 = Fi 2 / Fm 9), while the absence of an absolute superlative in *-issimo* would seem to be more significant in the case of *significativo* (Fi 0 / Fm 1,163) or *caratteristico* (Fi 0 / Fm 325) than in that of *letterario* (Fi 0 / Fm 10).[4] The greater Fm, the smaller the chance that the absence of *-issimo* might be casual. Even as great a ratio as that of *significativo* (0 / 1,163), however, is not a garantee of the complete impossibility of the corresponding formation in *-issimo*: e.g. *significativissimo*, is indeed attested once from the unsuspicious pen of Livio Petrucci (see Serianni & Trifone 1994:51), and a later inspection of the Internet (17-06-02) permitted to spot three more examples. One must thus be extremely careful in calling definitively uncacceptable a certain formation in *-issimo*. Nevertheless, our probe has shown that there is an interesting correlation between frequency and acceptability.

4. The core group

One of the basic tenets of studies on untutored language acquisition is that learners essentially may only rely on positive evidence (see Sokolov & Snow 1994 for a good review of the literature). This is certainly also true of the acquisition of *-issimo* in Italian. No Italian child is ever told, neither at home nor at school, to avoid the absolute superlative *significativissimo*, and nevertheless most Italian speakers feel unhappy with this word. The source of this feeling must thus lie in a pattern abstracted from positive evidence, *viz.* the sum of formations in *-issimo* they have heard during their life-time.

One reasonable assumption is that the productive pattern is abstracted from those adjectives which show the highest Pi, i.e. from the most typical bases. Limiting our calculations to the 74 most frequently intensified adjectives (Fm + Fi > 1,000) and excluding six distorting cases (*santo* 'holy' [Pi = 316.71!]), *nuovo* 'new', *puro* 'pure', *notevole* 'remarkable', *vero* 'true' and *moderno* 'modern'), the avarage Pi of this group turns out to be 1.38, ranging from 5.92 (*prezioso* 'valuable, important') to 0.001 (*simile* 'similar'). Since, due to our sample, all the bases are highly frequent adjectives, the length of the bases (henceforth *Lb*) is, of course, also relatively low (average: 2.81 syllables). There is certainly no general prosodic restriction on the bases of *-issimo*, as *raccomandabilissimo* 'highly recommendable' and

similar formations show, but there are nevertheless interesting correlations between Pi and Lb, as we will see. Going through the list of typical bases,[5] one gets the impression that they all imply a high emotional involvement on the part of the speaker/writer (recall Wierzbicka's statement quoted above) and, in typical conditions of use, an unconditional commitment to the judgement expressed. In this respect, *-issimo* is similar to exclamative sentences which also combine intensification and emotivity: *Che bello*! 'How beautiful!' / *Com'è bello*! 'How beautiful!' ≈ *È bellissimo*! 'He/she/it's really beautiful!', etc. The existence and importance of this pragmatic restriction on the bases, which, unfortunately, is not readily measurable in as an objective and simple way as Lb, is already confirmed, it seems to me, by our bottom of the list, *viz. simile*, where 9,826 collocations of *molto simile* 'very similar' stand against only 13 cases of *similissimo* 'extremely similar'. 'Similar', in effect, is a predicate that is not uttered to perform a spontaneous and highly emotional judgement, but presupposes a situation where the speaker, in a considered way, compares two entities in order to assess their degree of similarity. As the high frequency of *molto simile* shows, similarity is often found to exist to a high degree, but since the judgement does not involve a high degree of emotion, speakers only very rarely use the suffix *-issimo* (or exclamative sentences). A second lesson that may already be learned from this case, together with the *significativissimo* case discussed above, is that the restrictions on the use of *-issimo* are not of an all-or-nothing type but of a gradual nature, which, by the way, supports the hypothesis that the fundamental restriction is of a semantic-pragmatic nature.

5. Identification and explanation of the lacunae

At the end of the last paragraph we have already moved away from the core group of typical bases to the other end of the scale where we find adjectives which do not readily combine with our suffix. It is by comparing the two extreme points of the scale that we may hope to gain more insight into the exact nature of the restrictions that govern the use of *-issimo*. Since the scope of the present paper is more of a methodological nature, it is not necessary here to repeat the detailed descriptions contained in Rainer (2003). Rather, it will suffice to chose eclectically some representative cases and generalizations that bear on the semantic-pragmatic hypothesis formulated in 4.

5.1. Synthetic comparatives

The group of Italian adjectives most clearly incompatible with *-issimo* are the eight synthetic comparatives in *-ore* listed in Table 1. The reason of this manifest incompatibility is that *molto* 'very' in front of comparatives does not express intensification, but a kind of quantification, while *-issimo* may not express this kind of quantification (note that many languages have special adverbs for comparatives: Eng. *very good vs much better*, Germ. *sehr gut vs viel besser*, Fr. *très bon vs beaucoup meilleur*, etc.). The oddness of synthetic comparative + *-issimo* thus follows straightforwardly from a characterisation of *-issimo* as an intensifying suffix. The two examples of *minorissimo* are only apparent exceptions, since *minore* is not used in the sense of 'smaller' but of 'minor' (*cf*. Eng. *very minor*). The same may be true of *superiorissimo* in the only example found: *Intellettualmente, l'uomo è superiorissimo* 'Intellectually, men are very superior'.

Table 1. Synthetic comparatives (data: 30-06-2000)

adjective	Fi/Fm	adjective	Fi/Fm
migliore 'better'	0/717	*peggiore* 'worse'	0/112
maggiore 'bigger'	0/2,072	*minore* 'smaller'	2/1,056
superiore 'superior'	1/1,911	*inferiore* 'inferior'	0/1,666
anteriore 'earlier'	0/42	*posteriore* 'later'	0/54

5.2. Bases in -ivo

One major difficulty with the use of the elicitation method in the study of single affixes is that informants tend to get confused about their own feeling for language after a couple of questions bearing on the same affix. Computers, on the contrary, can't get confused, so they will answer with objectivity any number of questions bearing on the same problem. In the following discussion, we will exploit this possibility of treating long series of repetitive data in order to describe the distribution of *-issimo* after bases in *-ivo*.

In Table 2, Italian adjectives in *-ivo* are cross-classified according to two parameters, *viz.* Pi and Fi: the more we move to the bottom, the lower the propensity for taking *-issimo*, and the more we move to the right, the lower the token frequency of the adjective in *-issimo*. The table shows, for example, that *cattivo* 'wicked' is a better base than *positivo* 'positive', *lascivo* 'licentious' or *passivo* 'passive' better

Table 2. Adjectives in *-ivo* and the suffix *-issimo* (data: June 2000)

Pi	Fi > 10	Fi > 1<10	Fi = 1
≥ 1	*cattivo* 'wicked'	*lascivo* 'licentious' *passivo* 'passive'	*permissivo* 'permissive'
≥ 0.1	*attivo* 'active' *esclusivo* 'exclusive' *sportivo* 'sporty'	*combattivo* 'combative' *oggettivo* 'objective'	*festivo* 'festive' *progressivo* 'progressive'
≥ 0.01	*positivo* 'positive'	*creativo* 'creative' *negativo* 'negative' *produttivo* 'productive' *selettivo* 'selective' *soggettivo* 'subjective'	*espansivo* 'extrovert' *nocivo* 'harmful' *offensivo* 'offensive' *primitivo* 'primitive'
≥0.001		*competitivo* 'competitive' *suggestivo* 'suggestive'	*espressivo* 'expressive' *impegnativo* 'demanding' *intuitivo* 'intuitive'

than *competitivo* 'competitive' or *suggestivo* 'charming'. Even more interesting for our purpose, of course, are those adjectives in *-ivo* – absent from Table 2 – for which there is no attested superlative in *-issimo* at all (Fi = 0). Here is the list, ordered according to Fm (Fm > 50): *significativo* 'significant' (1,239), *aggressivo* 'aggressive' (171), *innovativo* 'innovative' (164), *istruttivo* 'instructive' (140), *relativo* 'relative' (101), *approssimativo* 'rough' (94), *riduttivo* 'restrictive' (76), *costruttivo* 'constructive' (62), *incisivo* 'incisive' (58). According to our characterisation of the semantics and pragmatics of *-issimo*, these adjectives should have a low probability of being used in highly emotional judgements. On the whole, it would seem to me, this prediction is borne out. Scientists, for example, very often refer to a high degree of significance (*molto significativo* 'highly significant', after all, is attested 1,239 times!), but predications involving such a collocation generally are proffered with consideration, not in an emotional tone. Since the suppression of emotion is a general trait of scientific discourse, the correlation between low Pi and scientific register, which I had noted in my earlier work on *-issimo*, eventually turns out to be a side effect of our characterisation of the semantics and pragmatics of the suffix.

Sceptics might object that the reluctance of *significativo* to take *-issimo*, after all, could also be a consequence of its utter length (no less than 6 syllables). This is an objection that has to be taken seriously, because there is indeed a neat correlation between Lb and Pi

Table 3. Tetrasyllabic adjectives in *-ivo*, arranged according to Pi

Fm + Fi ≥ 50	(Fi/Fm = Pi)	Fm + Fi < 50	(Fi/Fm = Pi)
esclusivo 'exclusive'	(21/76 = 0.27)	*combattivo* 'combative'	(4/27 = 0.14)
creativo 'creative'	(4/79 = 0.05)	*oggettivo* 'objective'	(1/9 = 0.11)
soggettivo 'suggestive'	(3/64 = 0.04)	*progressivo* 'progressive'	(1/10 = 0.1)
negativo 'negative'	(5/182 = 0.02)	*offensivo* 'offensive'	(1/19 = 0.05)
positivo 'positive'	(30/1,045 = 0.02)	*espansivo* 'extrovert'	(1/21 = 0.04)
primitivo 'primitive'	(1/76 = 0.01)	*effettivo* 'effective'	(0/6 = 0.00)
produttivo 'productive'	(2/120 = 0.01)	*lucrativo* 'lucrative'	(0/6 = 0.00)
selettivo 'selective'	(2/113 = 0.01)	*difensivo* 'defensive'	(0/8 = 0.00)
espressivo 'expressive'	(1/438 = 0.002)	*esplosivo* 'explosive'	(0/8 = 0.00)
suggestivo 'suggestive'	(2/711 = 0.003)	*nutritivo* 'nourishing'	(0/8 = 0.00)
incisivo 'incisive'	(0/58 = 0.00)	*remissivo* 'compliant'	(0/8 = 0.00)
costruttivo 'constructive'	(0/62 = 0.00)	*discorsivo* 'chatty'	(0/9 = 0.00)
riduttivo 'restrictive'	(0/76 = 0.00)	*impulsivo* 'impulsive'	(0/10 = 0.00)
relativo 'relative'	(0/101 = 0.00)	*obiettivo* 'objective'	(0/10 = 0.00)
istruttivo 'instructive'	(0/140 = 0.00)	*attrattivo* 'attractive'	(0/11 = 0.00)
aggressivo 'aggressive'	(0/171 = 0.00)	*persuasivo* 'persuasive'	(0/11 = 0.00)
		tempestivo 'timely'	(0/12 = 0.00)
		distruttivo 'destructive'	(0/16 = 0.00)
		istintivo 'impulsive'	(0/17 = 0.00)
		riflessivo 'reflective'	(0/21 = 0.00)
		descrittivo 'descriptive'	(0/23 = 0.00)
		intensivo 'intensive'	(0/23 = 0.00)
		ricettivo 'receptive'	(0/24 = 0.00)
		sbrigativo 'hasty'	(0/24 = 0.00)
		protettivo 'protective'	(0/26 = 0.00)
		operativo 'operative'	(0/27 = 0.00)
		emotivo 'emotional'	(0/36 = 0.00)
		restrittivo 'restrictive'	(0/42 = 0.00)
		comprensivo 'foregiving'	(0/44 = 0.00)

among the adjectives in *-ivo* (and more generally): 2.81 (core group) < 3.25 (P ≥ 1) < 3.57 (Pi 0.1-0.9) < 3.90 (Pi 0.01-0.09) < 4.40 (Pi 0.001-0.009) < 4.55 (Pi = 0). It will thus be appropriate to keep Lb constant and see whether bases with high and low Pi still differ in the way our hypothesis predicts. In Table 3, all tetrasyllabic bases in *-ivo* are arranged according to Pi, in decreasing order (the left row contains the adjectives with Fm + Fi ≥ 50, the right one those with Fm + Fi < 50 and > 5). Even though we have no objective measure of "emotivity", I think we may say that the facts are compatible with the predictions of our semantico-pragmatic characterisation of *-issimo*.[6]

6. Conclusion

Our case study, it would seem to me, allows us to conclude that the Internet is indeed an important new research tool for studying

morphological productivity and especially the restrictions on patterns of word-formation.

A first advantage of the Internet is that, due to its enormous extension, the absence of a certain formation from the corpus is beginning to be interpretable as a reliable indicator of its low acceptability, at least with highly productive affixes like *-issimo*. With the steady growth of the Internet corpus itself, the reliability will still increase, especially in the case of low frequency words. The figures obtainable for *molto significativo* 'highly significant' on 27th October 2002 with the aid of Google, for example, are already more than six times (!) higher (7,120) than those gathered two years ago with Altavista (1,163).

At the same time, the Internet also shows that "Never say no" is a motto to be taken at heart in word-formation more than elsewhere. The enormous extension of the Internet corpus in fact enables us to find formations which are so rare or marginal that native speakers often tend to reject them in elicitation experiments. Thus, for example, Google also gives us, on the same day mentioned above, four cases of *significativissimo* 'extremely significant', a formation judged of dubious acceptability by most informants. The difference in frequency (Fi = 4 / Fm = 7,120), of course, remains highly significant and shows, that *significativissimo* may not be treated on a par with, say, *attivissimo* 'extremely active' (Fi = 6,330 / Fm = 6,610, Google, 27-10-02). A systematic investigation of the frequencies of all relevant items demonstrates something that morphologists have always known intuitively, *viz.* the gradual nature of most restrictions on patterns of word-formation, especially if they are pragmatic or semantic, as in the case of *-issimo*.

The Internet thus undoubtedly opens new perspectives for morphological research. It should not be forgotten, however, that it is not a representative sample of a language, since some types of texts are overrepresented, others underrepresented. But the same is true of other large corpora, especially newspaper and literary corpora. And since few languages are fortunate enough to possess large representative corpora like those existing for English, the Internet may constitute, for the time being, an invaluable substitute.

Address of the Author:

Wirtschaftsuniversität Wien, Institut für Romanische Sprachen, Augasse 9, 1090 Wien, Austria <franz.rainer@wu-wien.ac.at>

Note

[1] The present article is a summary of those aspects of Rainer (2003) that might be of interest to morphologists in general and not only to students of Italian.
[2] See Gaeta (2003).
[3] For a critical reaction to this observation, see Dressler & Merlini Barbaresi (1994:497).
[4] The numbers were gathered on 10th October 2000. Were not otherwise indicated, the data is from spring 2000. The search engine used was Altavista.
[5] It may suffice here to quote the first thirty most typical bases, with Pi ≥ 1: *prezioso* 'valuable, important', *famoso* 'famous', *caro* 'dear, expensive', *alto* 'high', *bello* 'beautiful', *antico* 'old', *breve* 'short', *giovane* 'young', *bravo* 'able, nice, etc.', *recente* 'recent', *grande* 'big', *dolce* 'sweet', *vasto* 'vast', *duro* 'hard', *grave* 'severe, etc.', *vivo* 'vivacious', *raro* 'rare', *lungo* 'long', *leggero* 'light', *rapido* 'fast', *valido* 'valid', *fine* 'fine', *piccolo* 'small', *ricco* 'rich', *noto* 'well-known', *felice* 'happy', *personale* 'personal', *simpatico* 'nice', *stretto* 'tight', *potente* 'strong'.
[6] A certain correlation could certainly be established again with the probability of finding the relevant adjective in exclamative sentences. *Che cattivo*! 'How wicked (he is, You are)!', e.g., is much better than *Che significativi che sono questi numeri*! 'How significant those numbers are!'. The only really puzzling fact, to me, is the low Pi of *aggressivo* 'aggressive', an adjective that would seem to be predestined to be used in emotive predications. The absence could be due to a euphonic reason, *viz.* the avoidance of the repetition of sibilants (*cf.* also the low acceptability of *prossimissimo* [4/320], *tossicissimo* [0/128], *prolississimo* [0/11]).

Bibliographical References

Dressler Wolfgang U. & Lavinia Merlini Barbaresi (1994), *Morphopragmatics. Diminutives and Intensifiers in Italian, German, and Other Languages*, Berlin etc., Mouton de Gruyter.

Gaeta Livio (2003) "Produttività morfologica verificata su *corpora*: il suffisso *-issimo*", in Rainer & Stein (2003: 43-60).

Gallaway Clare & Brian J. Richards, eds. (1994), *Input and interaction in language acquisition*, Cambridge, Cambridge University Press.

Rainer Franz (1983a), "Intensivierung im Italienischen", Salzburg: Institut für Romanistik (Salzburger romanistische Schriften VII).

Rainer Franz (1983b), "L'intensificazione di aggettivi mediante *-issimo*", in: Dardano Maurizio, Wolfgang U. Dressler & Gudrun Held (eds.), *Parallela. Akten des 2. Österreichisch-italienischen Linguistentreffens*, Rome 1982. Tübingen: Narr: 94-102.

Rainer Franz (2003) "Internet come strumento di lavoro per il morfologo: le restrizioni di *-issimo*", in Rainer & Stein (2003: 97-116).

Rainer Franz & Achim Stein (eds.) (2003), *I nuovi media come strumenti per la ricerca linguistica*, Frankfurt am Main etc., Lang.

Serianni Luca & Pietro Trifone (1994), *Storia della lingua italiana*, vol. III, Torino, Einaudi.

Sokolov Jeffrey L. & Catherine E. Snow (1994) "The changing role of negative evidence in theories of language", in Gallaway & Richards (1994:38-55).

WIERZBICKA Anna (1991) "Italian reduplication: its meaning and its cultural significance", in *id.*, *Cross-Cultural Pragmatics. The Semantics of Human Interaction*, Berlin etc., Mouton de Gruyter, 255-284.