Measures of fuzzy disarray in linguistic typology

Luca Bortolussi, Andrea Sgarro
Dept. of Mathematics and Informatics
University of Trieste
Trieste, Italia
luca@dmi.units.it, sgarro@units.it

Liviu P. Dinu
Facultatea de matematică
University of Bucharest
Bucharest, Romania
liviu.p.dinu@gmail.com

Abstract

We extend crisp measures of disarray to the vaguely defined context of natural languages, so as to tackle problems of linguistic typology related to the order of words. We deal both with short abstract structure of the type "subject verb object", and with texts in the original language and in a translation. Preliminary experimental results are provided.

Keywords: Measures of disarray, fuzzy permutations, rank distance, bubble distance, Spearman footrule.

1 Introduction

Every linguist would agree that words in Italian and Spanish have basically the same order, while the "ordinal structure" changes more and more as one moves from Italian to English, or to German, or even to Turkish, which is definitely faraway. While the order of words is recognised as an important typologic feature of languages, linguists have some trouble if they want to deal with it in a "precise" way. Actually, the only ordinal structure which is dealt with at length in linguistic typology is rather abstract and is related to the position of subject S, verb V and object O in a simple *unmarked* sentence like *mater filium amat* and its translations (unmarked means that one should avoid those changes which may serve to connote "emotionally" the utterance, or which one uses in subordinate clauses, and the like; actually, Latin has a rather free word order). Unlike Latin, which has SOV, modern European languages prefer the order SVO, as in Italian, French, Spanish, English, German, Russian, or also Hebrew ('ivrit) etc. but Turkish, Basque and modern Persian (Fārsi) have SOV like Latin, while Irish (Gaeilge) has VSO. If one counts the number of *twiddles* between adjacent positions, the *ordinal distance* between the Latin SOV and the Italian SVO, or between the Italian SVO and the Irish VSO, is equal to one, while it goes up to two for Latin compared directly to Irish. This allows one to make a "precise" statement like: thinking of subject, object and verb in their unmarked order, the ordinal distance between Italian and Irish is strictly less that the ordinal distance between Irish and Latin. As soon as one proceeds to more ambitious tasks (think of a long text with a translation) one has to be contented with "feelings", say the trouble taken by a simultaneous interpreter to do her/his job. The point we wish to make in this paper is that one can deal with ordinal differences in a precise mathematical way, but that the mathematical tools one needs are not those of

crisp mathematics, but rather those of its soft extensions as fuzzy logic, due to the fact that natural languages have a "soft nature" which should not be unnaturally forced into the crisp fetters of traditional mathematics.

Let us begin by insisting on abstract structures as the one above, and let us introduce the negation N: *la madre non ama il figlio, the mother does not love the son, die Mutter liebt den Sohn nicht*, or, in Turkish, *ana oğlu sevmiyor*. If one takes a crisp attitude, one ends up with SNVO in Italian and English, SVON in German and SOVN in Turkish. Linguists might complain, and say that in English the verb "wraps up" the negation (because of *does*), and even more clearly so in Turkish: *sev-m[e]-iyor*, where *sev* carries the meaning and *iyor* the functionality, compare with *seviyor*, *[she] loves*). The linguist might insist that also the orders SVNO in English and SONV in Turkish are "true at some degree", and so she/he might re-discover fuzzy logic, or at least might have a hunch of it. The problem of fuzziness becomes much more dramatic when one moves from simple abstract structures to long and complex "real" texts, cf. Section 4 below.

What we need is to be able to measure the degree of *disarray* between two permutations of n distinct objects, i.e. the integers from 1 to n. Two *crisp* answers are the *bubble distance* (so called from the well-known bubble algorithm, a suboptimal algorithm for sorting) and the *rank distance*, already used in computational linguistics [1] and tightly related to *Spearman footrule*, as covered in [2], which is the basic reference on measures of disarray. However, we shall have to introduce a *fuzzy* generalisation of the notion of a permutation; all of this is covered in Sections 2 and 3. Some preliminary experimental results, both on abstract structures and on texts, are presented in Section 4, whose last paragraph is devoted to perspectives and future work. We have tried to keep the main body of the paper accessible to a wider audience interested in linguistic applications, and so have relegated more technical material to the last two Sections 5 and 6.

When it comes to comparing the merits of measures of disarray based on the bubble distance vs. measures based on the rank distance, one has to recall that bubble distances are more accurate, while rank distances are computationally quicker. In Section 5 we hint at an alternative and promising fuzzy approach to bubble disarrays to be applied when dealing with short abstract structures. Instead, the rank distance as in Section 3 recommends itself when processing long texts; cf. Sections 4 and 5.

2 Ordinal distances: a reminder

Take a string x of n distinct "objects", whatever their nature, and a permutation[1] y of those n objects. If one needs to measure the degree of disarray of y with respect to x, possible answers are two ordinal distances called the *bubble distance* and the *rank distance*: the more its bubble or rank distance from x is high, the more its disarray w.r. to x; cf. [2]. The bubble distance $d_B(x, y)$ simply counts the minimal number of twiddles between adjacent positions one needs to take y back to x; it can be computed in *quadratic* time by means of the bubble algorithm for sorting (actually, by its definition, the bubble distance *is* the number

[1] A permutation can be always seen as a couple of *strings* which have the same composition; if the first string is somehow implicit (think of the first n integers in their natural order, or of n letters in lexicographic order) the permutation can be also seen as a *single* string. This convenient "ambiguity" will be used below.

of iterations the bubble algorithm goes through). Instead the rank distance $d_R(x, y)$ is defined as:

$$d_R(x, y) \;=\; \frac{1}{2} \sum_{\text{all } a} |i_x(a) - i_y(a)|$$

where $i_x(a)$ is the position occupied by object a in the string x. The reason for dividing[2] the sum by 2 is because we want that both distances have the same *unit*, assigning the value 1 to two strings which differ because of a single twiddle.

Example. Take x =ROMA and y =AMOR. Applying the bubbling algorithm, one has: AMOR $\rightarrow$ AMRO $\rightarrow$ ARMO $\rightarrow$ RAMO $\rightarrow$ RAOM $\rightarrow$ ROAM $\rightarrow$ ROMA, and so $d_B(x, y) = 6$. Instead $2d_R(x, y) = |i_x(\mathrm{R}) - i_y(\mathrm{R})| + |i_x(\mathrm{O}) - i_y(\mathrm{O})| + |i_x(\mathrm{M}) - i_y(\mathrm{M})| + |i_x(\mathrm{A}) - i_y(\mathrm{A})| = |1 - 4| + |2 - 3| + |3 - 2| + |4 - 1|$, and so $d_R(x, y) = 4$.

Apart from their metric[3] properties which are easily proved (e.g. the triangle inequality), we wish to stress at least two ordinal properties of these distances:

i) Invariance: if ϕ is a permutation, then $d(\phi(x), \phi(y)) = d(x, y)$.

A consequence of invariance is that, when the objects to permute are the integers from 1 to n, one may often assume to no restriction that x has the integers in their natural order, as we do now when stating property *ii)* (y_i is the i-th entry of y):

ii) Monotonicity: let x be the first n integers in their natural order, and let z be obtained from y by a single twiddle between $y_i < y_{i+1}$; then $d(x, z) \geq d(x, y)$.

More precisely, while in the case of the bubble distance, each such twiddle between $y_i < y_{i+1}$ contributes 1 to the distance, in the case of the rank distance the contribution is either 1 or 0. It is 0 (in a way, the twiddle is "missed") when it occurs "remotely" enough from the original positions (cf. [2] for details). In practise, high distinct bubble distances may correspond to a single "squashed" value of the rank distance. If the rank distance may appear sloppier than the bubble distance on large distances, it has a paramount advantage when processing large-scale data, as those mentioned in Section 4: namely, its computation takes only *linear* time [3].

The *maximal* values and the *expected* values follow in function of the string length n; we shall use the term *random* distance, rather than expected distance, because this is better understood by linguists: the situation of greatest confusion corresponds to the random value, which one expects to find when y is obtained by shuffling the components of x totally at random. Instead, the maximal values are obtained when y is the mirror image of x, which implies that the ordinal structures of x and y are deeply related (below angular brackets denote degree of truth):

	random	max
rank	$\frac{n^2-1}{6}$	$\frac{n^2-\langle n \text{ odd}\rangle}{4}$
bubble	$\frac{n^2-n}{4}$	$\frac{n^2-n}{2}$

The curious asymmetry for the rank random value, which lies at two thirds of the way, and not at half way, once more can be accounted for by the sloppiness of the rank distance on high distances.

In the case of large-scale data, one has often to compare distances $d(x, y)$ and $d(u, w)$ when the

[2] In the literature, usually one does not divide by 2. We stress that Spearman footrule as in [2] coincides with our rank distance only when x is made up by the integers from 1 to n in their natural order; this however is enough to recycle most results of [2], because of invariance (property *i*).

[3] In principle, a measure of disarray of y w.r. to x needs not even to be symmetric; we shall come back to this point in Section 4 when dealing with large-scale applications.

string length is not the same in the two cases. That is why in Section 4 we have normalised results on texts over the interval $[0, 100]$. In the case of the rank distance, always thinking of better readability, we have made use of a double normalisation: distances up to the random value have been mapped to the interval $[0, 50]$, while distances from the random value upwards have been mapped to $[50, 100]$; this way, the normalised value corresponding to maximum confusion is always 50.

3 Measures of fuzzy disarray

As we have already seen above with "abstract" structures, in linguistics one may have *splitting* and, conversely, *merging*: this implies that a string of distinct objects like $abcdef$, say, might permute to $cbabf[de]$: the object b is at the same time before and after a, and the two objects d and e have merged into the single object positioned after f; the square brackets mean that the objects inside them are not to be seen in the order as they are written down, but rather in an undefined order. Such a situation is compatible with several crisp permutations, in our case four, namely $cbafde$, $cabfde$, $cbafed$, $cabfed$.

Definition. A *fuzzy permutation* Y of a crisp string x is a uniform fuzzy singleton whose elements are crisp permutations of x.

Recall that a fuzzy singleton[4] is a fuzzy set whose fuzzy size $\sum_i \mu_i$ is constrained to be 1. So, because of uniformity, each permutation in Y will be assigned the same *degree of membership* $\mu_i = 1/h$, $h = |Y|$ being the size of the support of Y. A comment is to the point, since uniformity may seem to be unnecessarily restrictive: in practise, we end up dealing with a *non-specific object*, rather than a fuzzy set proper. A more general approach to fuzzy permutations is put forward in Section 6, but the "simplistic" approach taken in the body of the paper appears to be the most convenient for the linguistic applications we have in mind; in particular, it spares the linguist the necessity of specifying explicitly degrees of membership for the component crisp permutations, a stumbling block similar to the one encountered in Bayesian statistics when one has to specify prior probabilities.

Let us move to measures of disarray, and go back to Y as above. The crisp permutations which make up the fuzzy permutation Y have bubble distances 5, 4, 6, 5 from $x = abcdef$, and rank distances 4, 4, 4, 4, respectively. We do not want to deal with a distance which is a *fuzzy number* (cf. however Section 6) and so we shall *aggregate* the h crisp integer distances to a single crisp value, simply by resorting to averages, which are widely understood:

$$d(x, Y) = \sum_{1 \leq j \leq h} \mu_i \, d(x, y_j) = \\ = \frac{1}{|Y|} \sum_{1 \leq j \leq h} d(x, y_j) \qquad (1)$$

E.g. with x and Y as above, the component distances aggregate to a crisp bubble distance equal to 5 and to a rank distance equal to 4, as the linguist would expect.

Starting from a crisp string x is enough to deal with applications on large-scale data as those mentioned in Section 4, but in Section 5 we have

[4] We stress that one is not dealing with a *random* permutation, but rather with a given permutation which happens to be *ill-defined*. In the present context there is *no* probabilistic uncertainty, and not even *incomplete knowledge*: the *vagueness* is here intrinsic in the nature of the context, and has nothing to do with the ignorance of the agent, which, at least in principle and in part, should be removable as the agent acquires new information.

to cover also the more problematic case when we want to measure the disarray of Y w.r. to a string X whose ordering is itself fuzzy, as needed to deal with the "small" abstract structures of linguistic typology.

4 Experimental results and perspectives

We begin by an "abstract" structures, thinking of the Italian sentences *la madre non ama il figlio* SNVO, and its translations in French, English, German and Turkish, *la mère n'aime pas le fils* SNVNO, *the mother does not love the son* SVNVO, *die Mutter liebt den Sohn nicht* SVON, and *ama oğlu sevmiyor* SOVNV, respectively. In English and in Turkish the verb "wraps up" the negation, while in French it is the negation *ne ... pas* which wraps up the verb. The length is $n = 4$, for which the random values are 3 and 2.5, while the maximum values are 4 and 6. In the table entries are not doubled when $d_R = d_B$:

	Fr	En	Ge	Tu
It	0.5	0.5	2	2; 2.25
Fr		0.5	1.5	2; 2.5
En			1.5	2; 2.25
Ge				1.75; 2

When comparing French, English and Turkish, we need a more general (and definitely more problematic) approach to fuzzy permutations, as explained and commented upon in Section 5.

We have also made some very preliminary experiments on the order of words in real texts; it is precisely the difficulties encountered which have convinced us of the necessity to move to fuzzy maths. We have used the bilingual book [4] for German readers, taking paragraphs at random from short stories in 6 different languages, and computing distances starting from the German translations, where we have singled out the variable parts of speech plus the negation *nicht*. So, the measures of disarray we got are non-symmetric disarrays from rather than symmetric disarrays between. We reproduce preliminary results normalised on the interval $[0, 100]$ (in the table stars denote normalisation). The reader is warned that nothing like statistical significance has to be expected: all this is rather sort of "warming up" before coping with really large texts, so as to avoid embarking on a long and difficult journey with an inappropriate baggage. We stress that web resources, inclusive of dictionaries and other tools available for automatic translations, allow one to sample really extensive data, so as to answer reliably questions like: how large is ordinal distance of Italian from Turkish, or of Turkish from Italian? Clearly, in these cases the linear complexity[5] of the rank distance turns out to be a decisive computational advantage.

from German	French	Spanish	Turkish
d_R^*	3.06	3.57	3.75
d_B^*	2.53	2.88	2.83

from German	Russian	English	Italian
d_R^*	3.33	2.08	3.90
d_B^*	2.26	1.65	3.39

Conclusion and open problems. The approach taken in the body of this paper is both easily understood by people outside the fuzzy community and computationally safe on large-scale data. However, to process effectively large-scale date one will need to extend and improve the fuzzy tools we have put forward in this paper (e.g. we have not covered deletions and insertions), beside

[5] With respect to usual rank distances, no increase in complexity occurs as soon as one assumes that mergings and splittings are *bounded*, in the sense that a letter is repeated at most K times, and at most K symbols merge, and this is reasonable in our context.

specifying in detail fast algorithms for the rank distance. As for short abstract structures, we are currently working on an alternative fuzzy version of the "more precise" bubble distance, extended in the spirit of the edit distance; cf. Section 6.

5 Double fuzziness

In Section 4, when X and Y are both fuzzy we have simply weighted the distances $d(x_i, y_j)$ by means of the uniform weights $1/hk$; in practise, XY is assumed to be a uniform fuzzy singleton of couples of permutations, $|X| = h$, $|Y| = k$, $1 \leq i \leq h$, $1 \leq j \leq k$. However, this has some serious drawbacks. The first is that one obtains positive self-distances $d(X, X)$ as soon as X is strictly fuzzy. This is not new in a fuzzy setting, since after all the fuzzy Hamming distance[6] has the same drawback. Rather, a more worrying fact is the following: both VNV, as in English, and NVN, as in French, give rise to the same couples $\{NV, VN\}$, and so *information loss* has occurred, which is quite undesirable. To get rid of both drawbacks, we deem that one should turn to modified ordinal distances, such as to be applied *directly* to non-crisp orderings as those entailed by splitting and merging. In the case of the bubble distance we might give smaller "costs" to twiddles involving non-crisp positions, in the spirit of Levenstein's *edit distance*. E.g. SOVNV → SVONV → SVNOV → SVNVO, with twiddle costs equal to 1/2, 1, 1/2, respectively, would reduce the distance between Turkish and English to 2, a result which the linguist might find more natural than 2.5 as in the table. From French to English the new distance would remain 1: SNVNO → SNNVO → SNVO → SNVVO → SVNVO, with costs equal to 1/2, 0, 0, 1/2 for each arrow; we are (legitimately) assuming that splitting and merging of the same symbol has cost 0. When splitting and merging are limited to couples, as is the case here, it is easy to prove that, when one string is crisp, the edit-distance approach is equivalent to ours in Section 3 (use e.g. induction on the number of twiddles). More care needs to be taken in the general case, but this falls outside the scope of this paper.

Unfortunately, if the bubble distance is a special case of the edit distance, the rank distance *is not*, and so a similar way-out does not appear to be feasible (nor desirable, since edit distances have quadratic complexity). Now, the more appealing applications are precisely those to very long texts, where use of the linear-complexity rank distance recommends itself, but in this case one may proceed as we have done in Section 4, where double fuzziness may be shunned to the reasonable price of having to deal with "distances from" rather than "distances between".

6 An addendum on fuzzy permutations

With more generality than in the body of the paper, a fuzzy permutation of a crisp string x may be defined as a fuzzy singleton which is not bound to be uniform. As for aggregations, one may use formula (1) omitting the last side, and so obtain a weighted average of the crisp distances $d(x, y_i)$ with weights μ_i. We do not feel that one should relinquish the constraint that the fuzzy set is a fuzzy singleton (that we are dealing with *one* permutation, be it fuzzy), else one may end up getting counter-intuitive results, because a component permutation might matter more in an average that it does when it stands crisply by itself.

[6] Remarkably, the fuzzy Hamming distance had already been used as early as 1967 by Ž. Muljačić in computational linguistics [5]: a further argument that the mathematics of natural languages has to be fuzzy.

Actually, fuzzy permutations might be defined also in a different way: one might specify the degrees of truth of statements as "letter a permutes to letter b", and end up with a *fuzzy matrix* of degrees of truth. Let us work out an example to show the difference between the two approaches. We assume that one starts with a fuzzy permutation Y made up of h crisp permutations y_i of x, each with degree of membership μ_i, $1 \leq i \leq h$. As an example, take $n = h = 3$, Y uniform with support $\{abc; bac; acb; cba\}$. Now we try to convert Y to a fuzzy permutation Z defined by a matrix "in a natural way". In the matrix we set $\nu_{\alpha,\beta}$ equal to $\max_i \mu_i$ for any two letters α, β, where the maximum is taken over values of the index i for which y_i permutes α to β (in the example, rows and column headings are a, b, c). The idea is that α permutes to β iff there is at least one permutation y_i where this happens, either the first, or the second, ..., or the h-th; then to use the maximum operator for logical disjunctions. In the example one obtains a 3×3 matrix representing Z whose 9 entries are all equal to $1/4$. However, W uniform with support $\{cba; bca; cab; abc\}$ gives rise to exactly the same Z as does Y, and so undesirable *information loss* has occurred in the conversion.

We add a comment of the "fuzziness" of our distances. Let us have a crisp universe $\mathcal{A}$, a crisp distance d between the crisp objects of $\mathcal{A}$, which in our case are permutations. By applying the *extension principle*, one obtains a fuzzy distance $d(X, Y)$ between fuzzy sets of $\mathcal{A}$ which is a fuzzy number (cf. e.g. [6]), and which one might defuzzify by means of a suitable aggregator. For simplicity we take $X = x$ crisp, and so the formula reads:

$$\langle m \in d(x, Y)\rangle = \max_{d(x,y)=m} \langle y \in Y\rangle$$

(m is a crisp number, in our case a crisp integer, angular brackets denote degree of membership). Clearly, if there are two or more values y_i which give the same $m = d(x, y_i)$, we are at variance with the approach taken in Section 3 and Appendix A, where each of these values gives its own contribution separately.

Acknowledgements. We wish to thank projects PRIN "Large-scale development of certified mathematical proofs" and FIRB LIBi for financial support.

References

[1] L.P. Dinu. Rank distance with applications to similarity of natural languages. *Fundamenta Informaticae*, 64 1-4, pp. 135-149, 2005.

[2] P. Diaconis and R.L. Graham. Spearman footrule as a measure of disarray. *Journal of the Royal Statistical Society*, Series B, 39(2), pp. 262-268, 1977.

[3] L.P. Dinu and A. Sgarro. A low-complexity distance for DNA strings. *Fundamenta Informaticae*, 73-7, pp. 361-372, 2006.

[4] Liebe hat vielen Sprachen. dtv zweischprachig, Langewiesche-Brandt, 1998.

[5] Ž. Muljačić. Die Klassifikation der romanischen Sprachen. *Roman. Jahrbuch*, 1967.

[6] L. Bortolussi and A. Sgarro. Fuzzy codebooks for DNA Word Design. IPMU 2006, Paris, France,pp. 2784-2790.