A Model of the Distribution of the Distances of Alike Elements in Dialogical Communication

Alexander Mehler
Bielefeld University
Computational Linguistics & Text Technology
Alexander.Mehler@uni-bielefeld.de

***Abstract*—In this paper we describe a model of the distribution of alike elements in dialogical communication. Our starting point is the interactive approach to alignment which views priming as a basic mechanism of verbal interaction. As this mechanism predicts short distances of the occurrences of primes on the one hand primed units on the other we concentrate on the distance effect induced by priming. The present paper focuses on prerequisites of measuring this effect. Thus, it can be seen as a starting point of operationalising the notion of alignment in terms of information theory.**

I. Introduction

In this paper we compute the probability distribution of the distances of alike elements which have been produced by different interlocutors in dialogical communication. The background of this model is the notion of alignment in communication (Pickering and Garrod, 2004). This approach to dialogue postulates that representations at various (linguistic and non-linguistic) modes and levels are aligned based on two processes:

- First, by means of priming processes as short-term mechanisms of information percolation between different or within certain levels of cognitive representation. Amongst others, these levels include the lexical, syntactic and semantic level of linguistic representation.
- Second, by means of routinisation serving for the expectation driven control of dialogue unfolding — this mechanism relates to the interpersonal coupling of long-term intrapersonal processes.

In this paper we concentrate on the first of these processes, that is, on priming. A basic claim of the interactive alignment approach is that in dialogical communication an utterance by an interlocutor A that activates a certain part of an aligned situation model raises the probability of an occurrence of another utterance related to that part of the situation model (either by A or by the other interlocutor B) (Pickering and Garrod, 2004). If we project this bipartite approach (including some utterance layer and the situation model) onto a single layer (e.g. a linguistic level of representation) we might reformulate this idea by saying that the occurrence x of a unit of that layer raises the probability of an occurrence of a primed unit in a short distance to x. By relating the priming relation of linguistic units to distance relations of occurrences (i.e. tokens) we connect the alignment model with Skinner's hypothesis (Zörnig, 1984a). However, by focusing on interacting interlocutors we might reformulate this hypothesis as follows: *An utterance of agent A raises the probability that agent B will utter an alike item (whether elementary or complex) in short distance to the first utterance where being alike means to be primed.*

It is this hypothesis on which we focus in this paper. That is, we deal with the measurement of alignment in terms of short distances of alike elements of a certain layer of linguistic representation where this distance-related priming effect is seen to be the more effective the more aligned the corresponding interlocutors. In systems theoretical terms we are interested in the emergence of certain probability distributions of the relations of elements of a certain code among interacting agents where these relations get more and more structurally coupled (e.g. more and more similar) during ongoing interactions. To the best of our knowledge this has not been investigated in information theory so far. However, there is the highly interesting and inspiring work of Zörnig (1984a,b) who has already studied the probability distribution of the distances of alike elements in written communication, that is, in monologues output by, so to speak, single interlocutors. In this paper we heavily build on this model and extend it in order to grasp interactions of two interlocutors in dialogical communication. However, we concentrate on finding a

probability distribution of the distances of alike elements generated by different agents in dialogical communication. Thus, we leave the extension of this model in terms of a measurement operation of alignment to future work.

II. Distances of Alike Elements in Dialogical Communication

A. Modeling Bilayer Strings and Sequences

Let $\mathbb{L} = \{a_1, \ldots, a_p\}$, $\mathbb{F} = \{b_1, \ldots, b_q\}$ be two sets and $t : \mathbb{F} \to \mathbb{L}$ a total function where $t(b) = a \in \mathbb{L}$ is called *type* of *form* $b \in \mathbb{F}$ — we alternatively call $\mathbb{L}$ the *lexicon* of all types manifested by at least one form in $\mathbb{F}$. By analogy to Zörnig (1984a) we assume that types and forms are identified by their index, that is, $\mathbb{L} = \{1, \ldots, p\}$ and $\mathbb{F} = \{1, \ldots, q\}$.[1] Now, let $\langle \mathbb{F}^\star, \circ, \epsilon \rangle$ and $\langle \mathbb{L}^\star, \circ, \epsilon \rangle$ be the free monoid over $\mathbb{F}$ and $\mathbb{L}$, respectively, based on the concatenation function $\circ$. Any $X = i_1 \circ \ldots \circ i_l \in \mathbb{F}^\star$, $i_j \neq \epsilon, j \in \{1, \ldots, l\}$, is called *string* and any $Y = t(i_1) \circ \ldots \circ t(i_l) \in \mathbb{L}^\star$ is called *sequence* both of length l. Next, we extend t as a function over strings. That is, for any string $X = i_1 \circ \ldots \circ i_l \in \mathbb{F}^\star$, $t(X) = t(i_1) \circ \ldots \circ t(i_l) \in \mathbb{L}^\star$ is a sequence. Now, we define *tokens* by their string and sequence position, respectively. That is, i_j and $t(i_j)$ is the jth token within string X and sequence $t(X)$, respectively. In order to refer to the tokens of a string $X = i_1 \circ \ldots \circ i_l$ or a corresponding sequence $Y = t(X)$ we define two functions $\mathsf{t}_X : \{1, \ldots, l\} \to \{i_1, \ldots, i_l\}$ with $\mathsf{t}_X(j) = i_j$ and $\mathsf{t}_Y : \{1, \ldots, l\} \to \{t(i_1), \ldots, t(i_l)\}$ with $\mathsf{t}_Y(j) = t(i_j)$. We alternatively call $\{1, \ldots, l\}$ the set of tokens of X and Y, respectively. Note that for two tokens $n, m \in \{1 \ldots, l\}$, $n \neq m$, it may be that $\mathsf{t}_X(n) = \mathsf{t}_X(m)$ (the same form) or $\mathsf{t}_{t(X)}(n) = \mathsf{t}_{t(X)}(m)$ (the same type).[2] Since t is a function we have: $\mathsf{t}_X(n) = \mathsf{t}_X(m) \Rightarrow \mathsf{t}_{t(X)}(n) = \mathsf{t}_{t(X)}(m)$.

Now suppose two strings $X_A = i_1 \circ \ldots \circ i_{n_A}$ and $X_B = j_1 \circ \ldots \circ j_{n_B}$ each representing an agent-specific layer of the same mode (e.g. gesture, lexis, syntax etc.) of a multilayer model of multimodal communication between two agents A and B. In order to map this bilayer scenario onto a unilayer model we suppose a string $X_{AB} = k_1 \circ \ldots \circ k_{n_A+n_B}$ called *linearisation* of X_A and X_B such that there exists a partition of the set $\{1, \ldots, n_A + n_B\}$ of tokens into A', $|A'| = n_A$, and B', $|B'| = n_B$, together with two bijections $\mathsf{pr}_A : A' \to \{1, \ldots, n_A\}$ and $\mathsf{pr}_B : B' \to \{1, \ldots, n_B\}$. This allows us to specify each token $i \in \{1, \ldots, n_A + n_B\}$ by the agent who has produced it as done by the function $\mathsf{ag}\colon \{1, \ldots, n_A + n_B\} \to \{A, B\}$ with $\mathsf{ag}(i) = A$ iff $i \in A'$ and $\mathsf{ag}(i) = B$ iff $i \in B'$.

Note that because of the existence of overlapping layers generating a linearisation is not trivial. However, as we concentrate on a formal model we disregard this task. Next, each type $i \in \mathbb{L}$ is seen to occur exactly $m_i = m_{i_A} + m_{i_B}$ times in X_{AB}, m_{i_A} times in X_A and m_{i_B} times in X_B such that

$$n = n_A + n_B = \sum\nolimits_{i=1}^{p} m_i = \sum\nolimits_{i=1}^{p} m_{i_A} + m_{i_B} \quad (1)$$

is the length of the linearisation X_{AB}.

As we do not count distances of forms, but of types we focus on sequences $t(X) \in \mathbb{L}^\star$ derived from strings $X \in \mathbb{F}^\star$. By analogy to Zörnig (1984a) but with a focus on dialogical communication we denote by $\mathbb{Y}_{m_1,\ldots,m_p}$ the set of all permutations of all types such that according to Equation 1 each type $i \in \mathbb{L}$ occurs exactly m_i times in each sequence $Y \in \mathbb{Y}_{m_1,\ldots,m_p}$. Now, we are on the level of Zörnig's model, but with the difference that Y results from linearising strings generated by different agents. In order to collect all notions introduced so far by a single definition we define *bilayer*s as tuples

$$\begin{aligned} \mathbb{L}_{A,B} \quad = \quad & (\mathbb{L}, \mathbb{F}, t, X_A, X_B, X_{AB}, \mathbb{Y}_{m_1,\ldots,m_p}, \\ & \{\mathbb{Y}_{r_{A|B}} \,|\, r \in \mathbb{L}\}, \\ & \{r_{A|B}\colon \mathbb{Y}_{m_1,\ldots,m_p} \to \mathbb{Y}_{r_{A|B}} \,|\, r \in \mathbb{L}\}, \\ & \mathsf{t}_{X_{AB}}, \mathsf{pr}_A, \mathsf{pr}_B, \mathsf{ag}) \end{aligned} \quad (2)$$

such that X_{AB} linearises X_A and X_B of length n_A and n_B, respectively, and for each $r \in \mathbb{L}$

$$\mathbb{Y}_{r_{A|B}} = \mathbb{Y}_{m_1,\ldots,m_{r-1},m_{r_A},m_{r_B},m_{r+1},\ldots,m_p} \quad (3)$$

is equal to $\mathbb{Y}_{m_1,\ldots,m_p}$ except that the m_{r_A} tokens of r generated by A and the m_{r_B} tokens of r generated by B are viewed as being of different type r_A and r_B, respectively. For a given $Y \in \mathbb{Y}_{m_1,\ldots,m_p}$ and type $r \in \mathbb{L}$ we use the function $r_{A|B} : \mathbb{Y}_{m_1,\ldots,m_p} \to \mathbb{Y}_{r_{A|B}}$ with $r_{A|B}(Y) = Y_{r_{A|B}} \in \mathbb{Y}_{r_{A|B}}$ for $Y \in \mathbb{Y}_{m_1,\ldots,m_p}$ — once more, $Y_{r_{A|B}}$ is identical with Y except that the m_{i_A} tokens $1 \leq i \leq n$, $t(i) = r$, for which $\mathsf{ag}(i) = A$ and the m_{r_B} tokens $1 \leq i \leq n$ of the same type for which $\mathsf{ag}(i) = B$ are seen to be of different type r_A and r_B, respectively.

Now, let a bilayer $\mathbb{L}_{A,B}$ be given and $i, j \in \{1, \ldots, n_A + n_B\}$ be two tokens of $Y_{AB} = k_1 \circ \ldots \circ k_{n_A+n_B}$ such that $1 \leq i < j \leq n = n_A + n_B$. Then we

[1] Note that Zörnig (1984a) does not distinguish types (or basic forms) and their instances. Note also that types and tokens might be complex or compound in terms of data oriented parsing (Bod et al., 2003).

[2] In terms of lexical items, the elements of $\mathbb{F}$ are word forms of lemmata in $\mathbb{L}$ while string positions denote lexical tokens.

define two distance functions with corresponding Boolean functions where Equation 4 and 6 the first and third are taken from Zörnig's model (Eq. 4 is only slightly adapted):

$$\delta_Y(k_\mu,k_\nu) = \begin{cases} \nu-\mu-1 \,:\, k_\mu=k_\nu \wedge \\ \qquad 1\le\mu<\nu\le n \\ \qquad -1 \,:\, otherwise \end{cases} \in\{-1,0,\ldots,n-2\} \quad (4)$$

$$\langle\delta\rangle_Y(k_\mu,k_\nu) = \begin{cases} \delta_Y(k_\mu,k_\nu) \,:\, \mathsf{ag}(\mu)\neq\mathsf{ag}(\nu) \\ \qquad -1 \,:\, otherwise \end{cases} \in\{-1,0,\ldots,n-2\} \quad (5)$$

$$\beta^{[r]}_{\mu,\nu}(Y) = \begin{cases} 1 \,:\, k_\mu=k_\nu=r \\ 0 \,:\, otherwise \end{cases} \quad (6)$$

$$\langle\beta\rangle^{[r]}_{\mu,\nu}(Y) = \begin{cases} 1 \,:\, k_\mu=k_\nu=r \wedge \\ \quad \mathsf{ag}(\mu)\neq\mathsf{ag}(\nu) \\ 0 \,:\, otherwise \end{cases} \quad (7)$$

Look at the following string in order to see how these functions work:

1	2	3	4	5	6	7
$\downarrow_{Y_{AB}}$	$\downarrow_{Y_{AB}}$	$\downarrow_{Y_{AB}}$	$\downarrow_{Y_{AB}}$	$\downarrow_{Y_{AB}}$	$\downarrow_{Y_{AB}}$	$\downarrow_{Y_{AB}}$
1 $\circ$	2 $\circ$	1 $\circ$	2 $\circ$	3 $\circ$	2 $\circ$	1
$\downarrow_{\mathsf{ag}}$	$\downarrow_{\mathsf{ag}}$	$\downarrow_{\mathsf{ag}}$	$\downarrow_{\mathsf{ag}}$	$\downarrow_{\mathsf{ag}}$	$\downarrow_{\mathsf{ag}}$	$\downarrow_{\mathsf{ag}}$
A	B	A	A	B	A	B

In this case we have $Y_{AB} = 1\circ 2\circ 1\circ 2\circ 3\circ 2\circ 1$, $k_1 = k_3 = k_7 = 1$, $k_2 = k_4 = k_6 = 2$, $k_5 = 3$. Further, $\delta_{Y_{AB}}(k_1,k_7) = 5$, $\delta_{Y_{AB}}(k_7,k_1) = \delta_{Y_{AB}}(k_2,k_7) = -1$ and $\delta_{Y_{AB}}(k_4,k_6) = 1$. Next, we see that $\beta^{[1]}_{1,7}(Y_{AB}) = \beta^{[1]}_{7,1}(Y_{AB}) = \beta^{[2]}_{4,6}(Y_{AB}) = 1$ and $\beta^{[1]}_{2,7}(Y_{AB}) = \beta^{[2]}_{2,7}(Y_{AB}) = 0$. Finally, $\langle\delta\rangle_{Y_{AB}}(k_1,k_7) = 5$, $\langle\delta\rangle_{Y_{AB}}(k_7,k_1) = \langle\delta\rangle_{Y_{AB}}(k_2,k_7) = \langle\delta\rangle_{Y_{AB}}(k_4,k_6) = -1$, $\langle\beta\rangle^{[1]}_{1,7}(Y_{AB}) = \langle\beta\rangle^{[1]}_{7,1}(Y_{AB}) = 1$ but $\langle\beta\rangle^{[2]}_{4,6}(Y_{AB}) = 0$.

Obviously, $\langle\delta\rangle_Y(k_\mu,k_\nu)$ counts only distances ≥ 0 of identical types which were produced by different agents, that is, which stem from different component layers of the linearisation X_{AB}. Now, we are in a position to redefine and extend Zörnig's model as follows:

1) $d^{[r]}_i(Y)$ is the number of all distances i (henceforth called i-distances) of pairs of identical types in sequence $Y\in\mathbb{Y}_{m_1,\ldots,m_p}$:

$$\begin{aligned} d^{[r]}_i(Y) &= |\{(k_\mu,k_\nu)\,|\,k_\mu=k_\nu=r\in\mathbb{L}\ \wedge \\ &\quad \delta_Y(k_\mu,k_\nu)=i\geq 0\}| \\ &= \sum_{\mu=1}^{n-i-1}\beta^{[r]}_{\mu,\mu+i+1}(Y) \end{aligned} \quad (8)$$

Analogously, $\langle d\rangle^{[r]}_i(Y)$ is the number of all i-distances of pairs of identical elements in $Y\in\mathbb{Y}_{m_1,\ldots,m_p}$ which were produced by different agents:

$$\begin{aligned} \langle d\rangle^{[r]}_i(Y) &= |\{(k_\mu,k_\nu)\,|\,k_\mu=k_\nu=r\in\mathbb{L}\ \wedge \\ &\quad \langle\delta\rangle_Y(k_\mu,k_\nu)=i\geq 0\}| \\ &= \sum_{\mu=1}^{n-i-1}\langle\beta\rangle^{[r]}_{\mu,\mu+i+1}(Y) \end{aligned} \quad (9)$$

2) The number of all i-distances between r-types in all sequences $Y\in\mathbb{Y}_{m_1,\ldots,m_p}$ is defined as:

$$D^{[r]}_i(\mathbb{Y}_{m_1,\ldots,m_p}) = \sum_{Y\in\mathbb{Y}_{m_1,\ldots,m_p}} d^{[r]}_i(Y) \quad (10)$$

Analogously, the number of all i-distances between r-types in all sequences $Y\in\mathbb{Y}_{m_1,\ldots,m_p}$ produced by different agents is defined as:

$$\langle D\rangle^{[r]}_i(\mathbb{Y}_{m_1,\ldots,m_p}) = \sum_{Y\in\mathbb{Y}_{m_1,\ldots,m_p}} \langle d\rangle^{[r]}_i(Y) \quad (11)$$

3) The number of all i-distances of pairs of identical types of whatever kind in all sequences $Y\in\mathbb{Y}_{m_1,\ldots,m_p}$ is defined as:

$$D_i(\mathbb{Y}_{m_1,\ldots,m_p}) = \sum_{r=1}^{p} D^{[r]}_i(\mathbb{Y}_{m_1,\ldots,m_p}) \quad (12)$$

Analogously, the number of all i-distances of pairs of identical types of whatever kind produced by different agents in all sequences $Y\in\mathbb{Y}_{m_1,\ldots,m_p}$ is defined as:

$$\langle D\rangle_i(\mathbb{Y}_{m_1,\ldots,m_p}) = \sum_{r=1}^{p}\langle D\rangle^{[r]}_i(\mathbb{Y}_{m_1,\ldots,m_p}) \quad (13)$$

Next, we have to derive calculable formulas of $\langle D\rangle^{[r]}_i(\mathbb{Y}_{m_1,\ldots,m_p})$ and $\langle D\rangle_i(\mathbb{Y}_{m_1,\ldots,m_p})$. This is done in Section II-B in order to prepare the calculation of the corresponding probability of i-distances of identical types generated by different agents according to the principle of chance (see Section II-C).

B. Calculating $\langle D\rangle^{[r]}_i(\mathbb{Y}_{m_1,\ldots,m_p})$ and $\langle D\rangle_i(\mathbb{Y}_{m_1,\ldots,m_p})$

Zörnig (1984a) proves the following theorem:

Theorem 1.

$$\begin{aligned} D^{[r]}_i(\mathbb{Y}_{m_1,\ldots,m_p}) &= (n-1-i)\frac{(n-2)!}{m_1!\ldots m_p!}m_r(m_r-1) \\ D_i(\mathbb{Y}_{m_1,\ldots,m_p}) &= (n-1-i)\frac{(n-2)!}{m_1!\ldots m_p!}(M-n) \\ &\quad \textit{where}\ \ M=m_1^2+m_2^2+\ldots+m_p^2 \\ D_i(\mathbb{Y}_{m_1,\ldots,m_p}) &= (n-1-i)D_{n-2}(\mathbb{Y}_{m_1,\ldots,m_p}) \end{aligned}$$

We complement and prove the following theorem:

Theorem 2.

$$\begin{aligned}
\langle D\rangle_i^{[r]}(\mathbb{Y}_{m_1,\ldots,m_p}) &= (n-1-i)\frac{(n-2)!}{m_1!\ldots m_p!}\cdot \\
&\quad (m_r(m_r-1)- \\
&\quad \frac{m_r!}{m_{r_A}!m_{r_B}!}(m_{r_A}^2+m_{r_B}^2-m_r)) \\
\langle D\rangle_i(\mathbb{Y}_{m_1,\ldots,m_p}) &= D_i(\mathbb{Y}_{m_1,\ldots,m_p})- \\
&\quad (n-1-i)\frac{(n-2)!}{m_1!\ldots m_p!}\cdot \\
&\quad \sum_{r=1}^{p}\frac{m_r!}{m_{r_A}!m_{r_B}!}(m_{r_A}^2+m_{r_B}^2-m_r) \\
\langle D\rangle_i(\mathbb{Y}_{m_1,\ldots,m_p}) &= (n-1-i)\langle D\rangle_{n-2}(\mathbb{Y}_{m_1,\ldots,m_p})
\end{aligned}$$

Proof. *The number $(d)_i^{[r]}(Y)$ of all i-distances of pairs of identical elements in $Y \in \mathbb{Y}_{m_1,\ldots,m_p}$ produced by the same agent (that is, either by A or by B) is given as:*

$$(d)_i^{[r]}(Y) = d_i^{[r]}(Y) - \langle d\rangle_i^{[r]}(Y) \quad (14)$$

Alternatively, we have:

$$\begin{aligned}
(d)_i^{[r]}(Y) &= d_i^{[r_A]}(r_{A|B}(Y)) + \\
&\quad d_i^{[r_B]}(r_{A|B}(Y))
\end{aligned} \quad (15)$$

Thus, we get:

$$\begin{aligned}
\langle d\rangle_i^{[r]}(Y) &= d_i^{[r]}(Y) - \\
&\quad d_i^{[r_A]}(r_{A|B}(Y)) - d_i^{[r_B]}(r_{A|B}(Y)) \\
&= \sum_{\mu=1}^{n-i-1}\beta_{\mu,\mu+i+1}^{[r]}(Y) - \\
&\quad \sum_{\mu=1}^{n-i-1}\beta_{\mu,\mu+i+1}^{[r_A]}(r_{A|B}(Y)) - \\
&\quad \sum_{\mu=1}^{n-i-1}\beta_{\mu,\mu+i+1}^{[r_B]}(r_{A|B}(Y))
\end{aligned} \quad (16)$$

Next, the number of all i-distances between r-types in all sequences $Y \in \mathbb{Y}_{m_1,\ldots,m_p}$ produced by the same agent is given as:

$$\begin{aligned}
(D)_i^{[r]}(\mathbb{Y}_{m_1,\ldots,m_p}) &= D_i^{[r]}(\mathbb{Y}_{m_1,\ldots,m_p}) - \\
&\quad \langle D\rangle_i^{[r]}(\mathbb{Y}_{m_1,\ldots,m_p})
\end{aligned} \quad (17)$$

and

$$\begin{aligned}
(D)_i^{[r]}(\mathbb{Y}_{m_1,\ldots,m_p}) &= D_i^{[r_A]}(\mathbb{Y}_{r_{A|B}}) + \\
&\quad D_i^{[r_B]}(\mathbb{Y}_{r_{A|B}})
\end{aligned} \quad (18)$$

so that

$$\begin{aligned}
\langle D\rangle_i^{[r]}(\mathbb{Y}_{m_1,\ldots,m_p}) &= D_i^{[r]}(\mathbb{Y}_{m_1,\ldots,m_p}) - \\
&\quad D_i^{[r_A]}(\mathbb{Y}_{r_{A|B}}) - \\
&\quad D_i^{[r_B]}(\mathbb{Y}_{r_{A|B}})
\end{aligned} \quad (19)$$

Now we can calculate according to Equation 19 and Theorem 1:

$$\begin{aligned}
&\langle D\rangle_i^{[r]}(\mathbb{Y}_{m_1,\ldots,m_p}) \\
&= D_i^{[r]}(\mathbb{Y}_{m_1,\ldots,m_p}) - D_i^{[r_A]}(\mathbb{Y}_{r_{A|B}}) - D_i^{[r_B]}(\mathbb{Y}_{r_{A|B}}) \\
&= (n-1-i)\frac{(n-2)!}{m_1!\ldots m_r!\ldots m_p!}m_r(m_r-1) - \\
&\quad (n-1-i)\frac{(n-2)!}{m_1!\ldots m_{r_A}!m_{r_B}!\ldots m_p!}m_{r_A}(m_{r_A}-1) - \\
&\quad (n-1-i)\frac{(n-2)!}{m_1!\ldots m_{r_A}!m_{r_B}!\ldots m_p!}m_{r_B}(m_{r_B}-1) \\
&= (n-1-i)\frac{(n-2)!}{m_1!\ldots m_p!}\cdot \\
&\quad \left(m_r(m_r-1)-\frac{m_r!}{m_{r_A}!m_{r_B}!}(m_{r_A}^2+m_{r_B}^2-m_r)\right)
\end{aligned}$$

Further:

$$\begin{aligned}
&\langle D\rangle_i(\mathbb{Y}_{m_1,\ldots,m_p}) \\
&= \sum_{r=1}^{p}\langle D\rangle_i^{[r]}(\mathbb{Y}_{m_1,\ldots,m_p}) \\
&= \sum_{r=1}^{p}(n-1-i)\frac{(n-2)!}{m_1!\ldots m_p!}\cdot \\
&\quad \left(m_r(m_r-1)-\frac{m_r!}{m_{r_A}!m_{r_B}!}(m_{r_A}^2+m_{r_B}^2-m_r)\right) \\
&= (n-1-i)\frac{(n-2)!}{m_1!\ldots m_p!}\left(\sum_{r=1}^{p}m_r^2-\sum_{r=1}^{p}m_r\right) - \\
&\quad (n-1-i)\frac{(n-2)!}{m_1!\ldots m_p!}\cdot \\
&\quad \left(\sum_{r=1}^{p}\frac{m_r!}{m_{r_A}!m_{r_B}!}(m_{r_A}^2+m_{r_B}^2-m_r)\right) \\
&= D_i(\mathbb{Y}_{m_1,\ldots,m_p}) - (n-1-i)\frac{(n-2)!}{m_1!\ldots m_p!}\cdot \\
&\quad \sum_{r=1}^{p}\frac{m_r!}{m_{r_A}!m_{r_B}!}(m_{r_A}^2+m_{r_B}^2-m_r)
\end{aligned}$$

Next, we have:

$$\begin{aligned}(n-1-i)\langle D\rangle_{n-2}(\mathbb{Y}_{m_1,\ldots,m_p}) &= (n-1-i)D_{n-2}(\mathbb{Y}_{m_1,\ldots,m_p}) - \\ &\quad (n-1-i)(n-1-n+2)\frac{(n-2)!}{m_1!\ldots m_p!}\cdot \\ &\quad \sum_{r=1}^{p}\frac{m_r!}{m_{r_A}!m_{r_B}!}(m_{r_A}^2+m_{r_B}^2-m_r) \\ &= D_i(\mathbb{Y}_{m_1,\ldots,m_p}) - \\ &\quad (n-1-i)\frac{(n-2)!}{m_1!\ldots m_p!}\cdot \\ &\quad \sum_{r=1}^{p}\frac{m_r!}{m_{r_A}!m_{r_B}!}(m_{r_A}^2+m_{r_B}^2-m_r) \\ &= \langle D\rangle_i(\mathbb{Y}_{m_1,\ldots,m_p})\end{aligned}$$

Based on Theorem 2 we are now in a position to calculate probabilities of the distances of alike elements generated by different interlocutors. This is done in the next section.

C. Building a Probability Distribution

Utilising the combinatorial model developed so far, we now calculate the probability p_i to randomly choose two tokens from a randomly chosen sequence $Y \in \mathbb{Y}_{m_1,\ldots,m_p}$ where the tokens are of the same type, have been generated by different agents and are exactly i unites distant from each other:

$$\begin{aligned}p_i &= \frac{\langle D\rangle_i(\mathbb{Y}_{m_1,\ldots,m_p})}{\sum_{j=0}^{n-2}\langle D\rangle_j(\mathbb{Y}_{m_1,\ldots,m_p})} \\ &= \frac{(n-1-i)\langle D\rangle_{n-2}(\mathbb{Y}_{m_1,\ldots,m_p})}{\langle D\rangle_{n-2}(\mathbb{Y}_{m_1,\ldots,m_p})\sum_{j=0}^{n-2}(n-1-j)} \\ &= \frac{2(n-1-i)}{n(n-1)} \qquad (20)\end{aligned}$$

Finally, we see that

$$\begin{aligned}\sum_{i=0}^{n-2}p_i &= \sum_{i=0}^{n-2}\frac{2(n-1-i)}{n(n-1)} \\ &= \frac{2}{n(n-1)}\sum_{i=0}^{n-2}(n-1-i) \\ &= 1 \qquad (21)\end{aligned}$$

so that we have to conclude that by means of Equation 20 we have build a correct probability distribution. Suppose now a bilayer $\mathbb{L}_{A,B} = (\mathbb{L},\ \mathbb{F},\ t,\ X_A,\ X_B,\ X_{AB},\ \mathbb{Y}_{m_1,\ldots,m_p}, \{\mathbb{Y}_{r_{A|B}} \mid r \in \mathbb{L}\},\ \{r_{A|B} \colon \mathbb{Y}_{m_1,\ldots,m_p} \to \mathbb{Y}_{r_{A|B}} \mid r \in \mathbb{L}\},\ \mathsf{t}_{X_{AB}},\ \mathsf{pr}_A,\ \mathsf{pr}_B,\ \mathsf{ag})$ is given such that

$$n = n_A + n_B = \sum\nolimits_{i=1}^{p} m_i = \sum\nolimits_{i=1}^{p} m_{i_A} + m_{i_B} \qquad (22)$$

We ask now for the expected number of i-distances given this distribution of the number of occurrences of $|\mathbb{L}|$ types. This expected value is given by

$$\begin{aligned}Np_i &= \left(\sum_{j=1}^{p}\binom{m_i}{2} - \sum_{j=1}^{p}\binom{m_{i_A}}{2} - \sum_{j=1}^{p}\binom{m_{i_B}}{2}\right)\cdot \\ &\quad \frac{2(n-1-i)}{n(n-1)} \qquad (23)\end{aligned}$$

where N is the number of all pairs of equal elements in a sequence $Y \in \mathbb{Y}_{m_1,\ldots,m_p}$ which have been generated by different agents A, B. This is where we depart from Zörnig's model as he considers a unilayer model generated by a single agent. Using Equation 23 we can compute for a sequence Y of length n and frequencies $m_1,\ldots,m_p$ of occurrences of types $1,\ldots,p$ in Y the theoretical distribution of the expected numbers of i-distances, $i = 0,\ldots,n-2$, according to chance. This is a central prerequisite of measuring the degree by which a given empirical distance distribution proves a significant distance effect in terms of aligned interlocutors.

III. Conclusion

In this paper we have built a probability distribution of the distances of alike elements produced by different interlocutors in dialogical communication. It can be used as a starting point of measuring the deviation of an empirical distribution of distances from the distribution of those distances which can be expected by the principle of chance. The present paper concentrated on formal, statistical aspects of this model. Its empirical testing will be part of future work.

Acknowledgment

Financial support of the German Research Foundation (DFG) through the SFB 673 *Alignment in Communication* (www.sfb673.org) — Project A3 *Dialogue Games and Group Dynamics* and Project X1 *Multimodal Alignment Corpora: Statistical Modeling and Information Management* — at Bielefeld University is gratefully acknowledged.

References

Altmann, G. (1985). On the dynamic approach to language. In Ballmer, T. T., editor, *Linguistic Dynamics. Discourses, Procedures and Evolution*, pages 181–189. De Gruyter, Berlin/New York.

Altmann, G. (1988). *Wiederholungen in Texten*. Brockmeyer, Bochum.

Bod, R., Scha, R., and Sima'an, K. (2003). A DOP model for phrase-structure trees. In Bod, R., Scha, R., and Sima'an, K., editors, *Data-Oriented Parsing*, pages 13–23. CSLI Publications, Stanford.

Garrod, S. and Anderson, A. (1987). Saying what you mean in dialogue: a study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–226.

Zörnig, P. (1984a). The distribution of the distance between like elements in a sequence I. In *Glottometrika 6*, pages 1–15. Brockmeyer, Bochum.

Zörnig, P. (1984b). The distribution of the distance between like elements in a sequence II. In *Glottometrika 7*, pages 1–14. Brockmeyer, Bochum.

Zörnig, P. (1987). A theory of distance between like elements in a sequence. In *Glottometrika 8*, pages 1–22. Brockmeyer, Bochum.