Referring Expressions and Communicative Success in Task-oriented Dialogues

Laura Aina, Natalia Philippova, Valentin Vogelmann, and **Raquel Fernández**
Institute for Logic, Language and Computation
University of Amsterdam
{laura.aina|natalia.philippova|valentin.vogelmann}@student.uva.nl
raquel.fernandez@uva.nl

Abstract

This paper studies lexical and structural properties of coreference chains in task-oriented dialogue and investigates their relationship with perceived and factual communicative success. In line with previous literature, our quantitative analysis shows that lexical entrainment is the most reliable predictor of task success, among the ones we compute. But also that there is a complex relationship between these factors – for example, neither high nor low, but rather intermediate levels of lexical alignment predict high perceived and factual success.

1 Introduction

The relationship between contextual information – broadly understood – and speakers' choices of referring expression is one of the most studied problems in both discourse and dialogue. In monological discourse, the main focus has been on contextual accessibility as a determinant of referring expression choice. For example, according to Ariel (1991), fully specified indefinite descriptions are used to refer to low accessibility entities – i.e., entities that are deemed to be completely unfamiliar to the audience – while definite descriptions, deictic expressions, and pronouns correspond to increasing levels of assumed accessibility (see, e.g., Orita et al. (2015) for a recent computational approach). In contrast, dialogue research has emphasised the fact that referring is a social act, drawing on evidence from the seminal work of Krauss and Weinheimer (1964), who showed that referring expressions get shorter when conversational partners provide ongoing feedback but not otherwise. In conversation, referring is not an autonomous act by the speaker who takes into account a generic audience, but rather a participatory act that requires coordinated actions from the addressee (Clark and Wilkes-Gibbs, 1986).

In this paper, we study the shape and dynamics of referring expressions in a classic reference-matching task between two dialogue participants who collaborate to build a puzzle. More concretely, we analyse lexical and structural properties of coreference chains (i.e., sequences of expressions with a common referent) and investigate the relationship between these properties and communicative success with respect to the referring task.

Several previous studies have considered the interdependence of speakers' linguistic choices and communicative success in task-oriented dialogue. Metzing and Brennan (2003) showed that participants took more time to find an object when their interlocutor suddenly switched referring expressions (e.g., by first referring to an object as *'the shiny cylinder'* and later as *'the silver pipe'*), thus breaking a conceptual pact (Brennan and Clark, 1996). Similarly, Nenkova et al. (2008) found that reuse of high-frequency words positively correlated with task success in a referential game. Reitter and Moore investigated syntactic and lexical repetition and showed that linguistic choices that reuse previously introduced material are more common in task-oriented dialogue and are reliable predictors of task success when repetition is present in the long-term (Reitter and Moore, 2007; Reitter and Moore, 2014). In contrast, Carbary and Tanenhaus (2011) and Foltz et al. (2015) found that only lexical alignment increased throughout the dialogue and positively affected task completion time.

Here we add to this line of research by making the following contributions: We develop three measures to quantitatively assess the dynamics of speakers' choices of referring expression, fo-

cusing on length, lexical repetition, and syntactic form matching. We apply these measures to a corpus of task-oriented dialogues in two languages, German and English, and provide a descriptive analysis of our findings. We then investigate the extent to which our measures are related to communicative success, distinguishing between *perceived* success and actual task success. Our results show that lexical repetition is the most reliable predictor of success in a non-trivial way: intermediate levels of lexical alignment (neither high nor low) predict high perceived and factual success. We end with a qualitative discussion of this and other findings of our study.

2 Dynamics of Referring Expressions

According to Clark and Wilkes-Gibbs (1986), the referring process often includes an *initiating* phase, a *refashioning* phase, and a *concluding* phase, during which referring expressions are grounded by the interlocutors. This can lead to lexical and structural entrainment (Brennan and Clark, 1996; Branigan et al., 2000), as well as to a simplification of the expressions over time (Krauss and Weinheimer, 1966), due to the establishment of consolidated antecedents (Ariel, 1991). In this section, we propose three simple measures to quantify these dynamics of referring expressions. The measures assume that referring expressions have been identified and are grouped into coreference chains, i.e., into chronologically ordered lists of expressions referring to the same entity.

Length Decrease. We are interested in a measure that allows us to quantitatively assess the degree to which the length of the expressions used to refer to a particular object declines over a certain timespan.

Let R^i be a coreference chain with referent i, i.e., a set of referring expressions used for the object i ordered chronologically in a given timespan. As an intermediate step, we define a measure of length drop for a referential expression $r_t^i \in R^i$ uttered at time step $t > 1$ and preceded by the set of expressions $R^i_{t'<t}$:

$$LenDrop(r_t^i) = \frac{L(R^i_{t'<t}) - len(r_t^i)}{L(R^i_{t'<t}) + len(r_t^i)}$$

where $len(x)$ is the number of tokens in an expression x and $L(X) = \mu(\{len(x)|x \in X\})$, that is the mean length of the expressions in coreference chain X. $LenDrop$ outputs a value in the range $[-1, 1]$: it is positive when the length of the target expression is shorter than the average length of the preceding expressions, negative when it is longer, and 0 when its equal to the average length.

We then operationalise the tendency towards length decrease within a coreference chain R^i as:

$$LenDecrease(R^i) = \mu(\{LenDrop(r_t^i) \mid r_t^i \in R^i_{t>1}\})$$

that is, as the average $LenDrop$ of each of the referring expressions in the chain. Since $LenDrop$ is undefined for the first phrase used to refer to i (it cannot be compared with any previous expressions), we compute the mean over all expressions except the first one, i.e., on $R^i_{t>1}$.

Lexical Alignment. Our second measure aims at capturing aspects of lexical entrainment, in particular the degree to which the choice of lexical items in a referring expression for object i involves words previously used in preceding expressions with the same referent. We define a function W that returns the set of content words in a set of referring expressions. We then compute the intersection of the content words $W(\{r_t^i\})$ of each expression $r_t^i \in R^i$ at time step $t > 1$ with the set of content words used in preceding expressions $W(R^i_{t'<t})$. We capture this information with the following ratio:

$$LexAlign(R^i) = \frac{\sum_{t=2} |W(\{r_t^i\}) \cap W(R^i_{t'<t})|}{|W(R^i_{t>1})|}$$

which expresses the relative frequency of choosing a content word in a referring expression that had already been used before to refer the same object. Similarly to $LenDecrease$, we define the ratio taking into account that the content words in the first phrase do not contribute to the overlap.

Form Alignment. Regarding syntactic form, our goal is to measure the extent to which speakers opt for constructing their referential expression using a type of phrase that had already been used before to refer to the same object. Let $syn(r_t^i)$ be the syntactic type of the referring expression r_t^i. Then, for $t > 1$:

$$FormAlign(R^i) = \frac{\sum_{t=2} [syn(r_t^i) \in F(R^i_{t'<t})]}{|R^i_{t>1}|}$$

where $F(X)$ is the set of syntactic types of the expressions in coreference chain X. $FormAlign$

hence measures the relative frequency of encountering a referring expression whose syntactic type has already been used before in a previous expression with the same referent. Again, when obtaining the denominator for normalization we do not consider the first referring expression.

An Example. To illustrate how these measures work, consider the following sequence of referring expressions used in this order to refer to a single puzzle piece:

(1) a. *a red piece on the left which looks like an elephant*
b. *the left piece next to the yellow one*
c. *the elephant*

The $LenDecrease$ of this coreference chain will be $\approx$ 0.4 (averaging over a $LenDrop$ of approximately 0.16 and 0.65 at the intermediate timesteps). *LexAlign* will be 0.5, as 3 over 6 content words had already been used before. *FormAlign* is instead 0.5 as only (1c) has the same syntactic type (definite noun phrase) as a preceding referring expression, in this case (1b).

3 Data

We use a subset of the human-human dialogues in the PentoRef corpus (Zarrieß et al., 2016), which consists of transcripts of conversations between two participants who can only communicate verbally and who work together to solve a Pentomino puzzle. In each dialogue, an *instruction giver* (IG), who has the full solution of the puzzle, directs an *instruction follower* (IF), who only has a board with an outline of the puzzle and the set of loose pieces. Their common goal is to get the IF to assemble the puzzle, which involves identifying pieces and locations on the board. The corpus is thus particularly suitable for studying human mechanisms related to choice of referring expressions and the relationship with task success.

The dialogues we leverage in the current experiment correspond to the control sections of the Push-to-Talk (Fernández et al., 2007) and Noise-NoNoise (Schlangen and Fernández, 2007) subcorpora of PentoRef — i.e., dialogues from experimental conditions with no manipulations.[1] The experimental setup for these control dialogues was identical, except for the fact that in the Noise-NoNoise experiment one puzzle piece was already placed on the right location on the board when the task started. In addition, the two sub-corpora differ in language: The participants were native English speakers in the Push-to-Talk corpus, while they are native German speakers in Noise-NoNoise. Since the measures we introduced in Section 2 are language-independent, we conduct our experiments on both sub-corpora. In the remainder of the paper, we will refer to the control sections of these sub-corpora as our experimental dataset and distinguish between the English and German section when needed. An overview of the dataset is provided in Table 1.

The corpus contains a range of annotations, including the identification of referential expressions together with the id of their referent (a piece or a location on the board) and their syntactic form (type of phrase, such as definite noun phrase, or pronominal phrase). The dialogues are divided into *moves*, where a move "covers all speech that deals with a particular piece, from the point when the players start to describe the piece [...] to the point when participants have agreed on the piece and its target location to their satisfaction and move on to the next piece" (Fernández et al., 2007). Each move is annotated for grounding status (i.e., the level of confidence of the participants on the placement of a piece) and for actual status on the board (i.e., the actual task success with respect to the puzzle solution).

Grounding status includes tags `confident`, `unconfident`, `on_hold`, and `reconfirm`. The first two indicate that the participants conclude a move placing a piece on the board with confidence or lack thereof, respectively. The tag `on_hold` indicates that the participants do not finish the placement of the piece before moving on to the next piece, while the tag `reconfirm` is used for moves where the participants go back to a piece that was already placed and leave it there. Board status includes tags `correct`, `wrong`, and `not_moved`. The first two options are about the success or not of a placement. The tag `not_moved` is used for moves where a piece has not been placed nor replaced (either because the move has been left unfinished or it consists of a reconfirmation without re-placement). Further details on the annotations are provided by Schlangen and Fernández (2008).

[1] Push-to-Talk includes an experimental condition where turn-taking is restricted, while Noise-NoNoise includes a condition where brown noise was added in real time at random points. We do not use these manipulated conditions in the present work; only the unrestricted dialogues.

	EN	DE	EN+DE
Dialogues	4	5	9
Utterances	1597	2764	4361
Moves	52	135	187
Utterances per move (μ)	30.7	20.5	23.3
Moves containing referring expressions	96%	99%	98%
Coreference chains per move (μ)	5.4	4.2	4.5
Coreference chains with length > 1	61%	59%	60%
Referring expressions per coreference chain (μ)	4.2	3.5	3.7

Table 1: Overview of the English (EN), German (DE), and combined (EN+DE) datasets

Table 3 gives an overview of the distribution of these communicative success tags for a subset of moves, as will be explained later on in Section 4.2.

4 Experimental Analysis

We now turn to applying our measures for quantifying the dynamics of referring expressions introduced in Section 2 to the dataset. We start by describing the results obtained for each measure, comparing them to a random baseline. We then move on to analysing the relationship between our measures and communicative success.

4.1 Descriptive Analysis of Dynamics

For each move in a Pentomino puzzle game, we compute the *LenDecrease*, *LexAlign* and *FormAlign* measures for each set of co-referring expressions mentioned in the timespan of a move. Since the three measures require to compare each expression to some previous ones and are hence undefined for singleton sets, we only take into account coreference chains that include more than one referring phrases. We evaluate the expressions uttered by the two participants collectively and also those uttered by only the IG or the IF, respectively. However, in our computations, we always consider the set of previous expressions to which the target expression is compared to be all the preceding referring expressions in the coreference chain, regardless of the speaker who uttered them.

To enable the interpretation of the results, we build a randomised baseline. For 100 iterations, we shuffle the order of expressions in the coreference chains spanning a dialogue, distribute them across moves respecting the original number of expressions in each move, and compute the measures on such a shuffled dataset. We then compare the distribution of the original data for each measure with the average of the shuffled distributions. This amounts to testing how crucial the chronological structure of the dialogue is for the investigated phenomena. It is worth pointing out that, given the limited vocabulary (constrained by the task at hand) and limited variety of phrase types, the baseline dialogues contain a considerable amount of local repetition despite the random shuffling. Any statistically significant values above the baseline will therefore be highly indicative of an effect. The statistics for each measure on our dataset and the random baseline can be seen in Table 2.

In order to level off the most relevant morphological difference between the English and German datasets, we make use of a compound splitter (Daiber et al., 2015) on the German referring expressions. This has effects on the length decrease and lexical alignment, as we treat compound components as separate tokens. For example:

(2) *in die* ***bauchseite*** $\rightsquigarrow$ *in die* ***bauch seite*** [*into the* ***side of the belly***]

After this pre-processing, as the phenomena we investigate and the methods we use are language-independent, we do not expect the English and German sections of the dataset to differ substantially. To test this, we compared the statistics obtained for each measure in the two languages, without finding any significant differences (Mann Whitney test $p > 0.01$ for all measures). Therefore, here and in the remainder of the paper, we report results on the combined dataset of English and German dialogues.

Length Decrease. We obtain an average of 0.08 *LenDecrease* across moves. The magnitude of such a decrease is significantly larger than the random baseline. We can thus conclude that in the

present dataset there is a general tendency for the length of referring expressions to decrease in the course of a dialogue, as attested in the literature. We did not find a significant difference between the length decrease of IG and IF.

Lexical Alignment. Our analysis of the reuse of lexical material yielded an average *LexAlign* result of 0.43, a value which is significantly higher than the random baseline. This confirms the well-known fact that in the course of task-oriented interactions speakers tend to progressively agree on a set of words to refer to a certain object. Although the mean value of *LexAlign* for the IF is higher than the one for the IG, there is no significant difference between the two distributions. However, while we found a significant difference between *LexAlign* of the IGs and the random baseline, the difference for the IFs is not statistically significant (possibly as a result of the high standard deviation $\sigma = 0.40$). Our analysis in the next section gives clues as to why this may be the case.

Form Alignment. We obtain an average of 0.75 *FormAlign*, a value that is significantly higher than the random baseline. This indicates that in on our dataset there is a tendency to match preceding syntactic forms within a coreference chain. IFs align significantly more than IGs in this case ($p < 0.01$). In addition, we found a positive correlation between Lexical and Form Alignment for all three levels of assessment – General, IG and IF (Spearman's $\rho \approx 0.20$; $p < 0.01$).

The results reported above seem to confirm some of the discourse tendencies attested for task-oriented dialogues in the literature. However, the high standard deviations of our measures point to a strong variability in the quantified phenomena across interactions. In the next section, we leverage precisely this variability to investigate whether patterns of use of referring expressions are informative with respect to communicative success.

4.2 Relationship to Communicative Success

As explained in Section 3, moves in the dataset are annotated for grounding status and world status. The former type of annotation codes perceived communicative success (i.e., the level of confidence of the participants) while the latter codes actual task success on the puzzle board. Here we analyse the possible interdependence between our measures and these levels of success. For this analysis, we consider as datapoints those moves that have communicative success tags and that contain referring expressions – this amounts to around 88% of the original dataset. Table 3 gives an overview of the communicative success of these moves. Note that grounding and board status interact significantly ($\chi^2 = 159.19,\ p < 0.01$), even if `not_moved`, `reconfirm` and `on_hold` are omitted ($\chi^2 = 8.38,\ p < 0.01$).

For each move in this subset, we calculate the mean and variance across the coreference chains in the move for each of our measures (*LenDecrease*, *LexAlign*, and *FormAlign*). We do this overall as well as for the IG and the IF independently. In addition, we compute the number of referring expressions per move. We exploit this information in two types of analyses: A comparison of distributions for different success levels and

		EN+DE		Random	
		μ	σ	μ	σ
LenDecrease					
General	**	0.08	0.25	0.02	0.03
IG	**	0.09	0.26	0.02	0.02
IF	**	0.09	0.23	0.03	0.03
LexAlign					
General	**	0.43	0.33	0.38	0.10
IG	**	0.44	0.36	0.38	0.08
IF		0.50	0.40	0.46	0.07
FormAlign					
General	**	0.75	0.32	0.70	0.24
IG	**	0.73	0.33	0.70	0.09
IF	**	0.82	0.32	0.74	0.08

Table 2: Mean and standard deviation for our three measures in the overall dataset and the random baseline; significance tested with Wilcoxon sum rank test (also known a MannWhitney), $^{**}p<0.01$

	`correct`	`not_moved`	`wrong`	Total
`confident`	84	1	10	95
`reconfirm`	0	24	1	25
`on_hold`	0	31	2	33
`unconfident`	5	1	5	11
Total	89	57	18	164

Table 3: Contingency table for moves that contain referring expressions and are annotated for grounding (rows) and board (columns) status.

a linear regression experiment where we estimate the probability of different levels of communicative success given our variables.

Comparison of Distributions. We start by testing whether our variables per move differ significantly across moves grouped by type of communicative success. We use the Wilcoxon rank sum statistical test (also known as Mann-Whitney) to check for significant effects and report *common language* (CL) effect size (McGraw and Wong, 1992) for the comparisons that are significant.

Regarding number of referring expressions per move, we find that moves tagged as `not_moved` – i.e., moves that do not lead to a piece being placed on the board – include significantly less referring expressions than other moves that led to identifying a piece and its location (mean number of referring expressions 12.77 vs. 25.6, $p < 0.001, \text{CL} = 0.69$). There are no significant differences in number of referring expression in `confident` vs. `unconfident` moves and `correct` vs. `wrong` moves.

As for *LexAlign*, we observe significant differences regarding grounding status, in particular on the behaviour of the IF: The IF reuses more lexical material in non-confident moves (`unconfident`, `on_hold`, `reconfirm`) than in confident moves (mean *LexAlign* 0.60 vs. 0.36, $p < 0.01, \text{CL} = 0.64$). In moves where the participants have confidently achieved grounding (according to their own perception, regardless of world status), there is less lexical alignment by the IF. Such a variable behaviour by the IF could explain the high standard deviation of $LexAlign$ reported in Table 2 and the lack of significant difference between the dataset and the random baseline distributions for the IF.

Concerning *LenDecrease* and *FormAlign*, we do not find any significant differences across communicative success levels when considering the mean values of these measures. There is simply a general tendency towards decreasing the length of referring expressions and towards reusing phrase types, as we have reported in Section 4.1. However, for *FormAlign* we observe significantly more variance in confidence successful moves for IG than for the IF (average variance of 0.028 vs. 0.015, $p < 0.01, \text{CL} = 0.40$). We do not find such a difference in variance for confident but wrong moves. The results of the regression confirm this effect and shed some light on the issue, as we discuss in the following paragraph.

Linear Regression Experiment. The observed differences between distributions grouped by type of success suggest that our measures do contain some information about the achieved level of communicative success. However, they do not yet specify the directionality of the relationship, i.e., which values of our measures are associated with which degrees of communicative success. To this end we perform a linear regression, where we estimate the expected value of success conditioned on the values of our variables. Since we are only interested in assessing directionality of the relationships and not primarily in accurate prediction, we opt for linear regression, the simplest possible model. This ensures a maximum of achievable interpretability of the estimated relationships and avoids further complicating the interpretation of the roles of our variables. Moreover, assuming communicative success to be a continuum approximated by the order of the categorical labels in Table 3, we can justify the assumptions made by linear regression models. Although the two types of success interact (as noted for Table 3), we construct separate models in order to shed light on the differences between the relationship of the type of success and our measures.

We consider the mean of each of our measures per participant role as described above as the main predictors in the regression models, but we also expect and include interactions of the mean with the following:

- Itself: This is equivalent to a quadratic transformation and allows for a non-monotone, specifically unimodal, relationship between the mean and the level of success. At least for some of the predictors, such a relationship may be more plausible, as extreme values of alignment may lead to similar probability of a certain success level, and vice-versa.
- Variance across coreference chains: High variance indicates low consistency of alignment across coreference chains in a given move, which may in turn signal communicative issues. We expect such issues to influence the achieved level of success at grounding and board levels.
- Number of referring expressions in the move: This is used as an indicator of the length of the communication needed to decide for an ac-

Grounding status (perceived success)				Board status (factual success)			
predictor	coeff.		SE	predictor	coeff.		SE
LenDecrease overall	-0.19	*	0.08	LexAlign IG	-0.38	***	0.1
LexAlign IF	-0.37	***	0.09	LexAlign IG ^{2}	-0.12	**	0.05
LexAlign overall ^{2}	0.22	*	0.09	LexAlign IG:num.exps	0.01	***	0.0
LexAlign IG ^{2}	-0.16		0.09	FormAlign:num.exps	-0.01	***	0.0
LexAlign IF ^{2}	-0.19	*	0.09	FormAlign IG mean:var	0.17	*	0.08
LenDecrease IF:num.exps	0.01	**	0.0				
LexAlign IG mean:var	0.26	*	0.14				

Table 4: Coefficients and standard error of the selected predictors of the linear regression models for perceived and factual success. Asterisks on the coefficients indicate their significance levels.

tion, and can intuitively be expected to affect both our measures and the resulting level of perceived and actual success.

The mean values of our measures together with these three types of interaction leave us with too many predictors to construct an informative regression model, and moreover we are interested in which of them have the highest predictive quality, so the first step is selection of predictors. We base this on persistence across different models, i.e., we select a predictor if the majority of models assigns it significant predictive value. In terms of the models we consider, we perform exhaustive search of all possible numbers and combinations of introduced mean values of our measures and their interactions with the three variables. For each predictor we then count its occurrence in models with high goodness-of-fit value. We select those predictors whose probability of occurring in a model with high goodness-of-fit is more than half.[2] Table 4 shows the predictors we selected for each type of success, where ^{2}, :var, and :num.exps refer to the introduced types of interaction, respectively.

Subsequently, we construct two regression models from the selected predictors of which we inspect the coefficients (found in Table 4) in order to assess the directionality with the respective type of success. In these models, all selected predictors are significant with the exception of squared *LexAlign* of the IG. The results are to be read as follows: If the predictor is not an interaction, a positive coefficient indicates higher probabilities of success of the respective type, and the opposite for negative coefficients. In the case of self-interaction, a negative coefficient translates to highest probability at intermediate value of the predictor, and at extreme values for a positive coefficient. As for the other two types of interaction, a positive coefficient indicates that the effect of our measure on the respective success type is higher if the interacting variable has a higher value, and vice-versa for negative.

Regarding the interpretation of these results, it is first to be emphasised that they are to be read with some caution: The goodness-of-fit of both models, and all models constructed for this analysis, is very low (respectively 0.15 and 0.12 adjusted r^2), and regression models become less reliable with lower goodness-of-fit values. On the other hand, this does not come as a surprise since it seems clear that success of both types does not only, or even mainly, depend on our measures. Looking at the coefficients in the model, it can however be stated with good confidence that it is not the case that high values of our measures lead to high levels of perceived and factual success (at least for those which were selected as predictors).

Furthermore, according to the models, the role of the IG prevails for factual success, while for perceived success, the IF's and overall alignment are more informative. See the next section for a discussion of this phenomenon. As already emergent in the previous analysis regarding the differences in distributions, *LexAlign* is clearly the most important of our measures when predicting communicative success. More specifically, *LexAlign* seems to have a non-monotone relationship with both types of success, i.e., neither high nor low, but intermediate levels of lexical alignment predict high perceived and factual success. The relationship of *LexAlign* with perceived success seems especially intricate: Note that due to the negative co-

[2]The procedure is based on the R language's `leap` function and validated by stepwise AIC selection.

efficient for *LexAlign* of the IF in the model, high probability of success is actually below intermediate and more towards no alignment. In contrast with the individual speaker roles, extreme values of overall *LexAlign* predict high success which alludes to the difference between individual and combined communicative effort.

Finally, the coefficients of the interaction terms where the measure itself was not selected as a predictor have no direct interpretation. At the same time, according to the way we model the relationship between our measures and success of both types, they are of high predictive value. We hence regard them as modelling artefacts and refrain from more detailed interpretation. As for the interaction of the mean value of *FormAlign* of the IG with its variance in factual success, it can however be stated that it confirms the observation made in the difference of variance in the previous section. This allows us to infer that *FormAlign* of the IG affects the probability of factual success to some degree, but the model does not give any information as to the directionality of the effect.

5 Discussion

The comparison of distributions and the results of the regression study suggest a complex relationship of the dynamics of referring expressions as captured by our measures with task success and grounding status. As already mentioned in the previous section, it is not simply the case that a higher degree of length decrease, lexical alignment and syntactic form matching leads to a higher probability of achieving success.

In line with previous literature, lexical alignment emerges as the most informative phenomenon and predictor, in particular when assessing separately the extent to which the IG or the IF reuse lexical material to compose their referring expressions. The effects are not symmetric for both participant roles: Lower lexical alignment by the IF leads to higher probability of confident grounding, while lower lexical alignment by the IG leads to higher probability of factual task success. In both cases, however, no lexical alignment also seems counter-productive.

Qualitative analysis of the dialogues indicates that low levels of lexical alignment by the IF in confident moves often are the result of grounding being achieved without the need to confirm or clarify, as in the following example:

(3) IG: *You know about the red cross?*
IF: *Yeah, I got it.*

In contrast, when confirmations and clarification requests are needed to achieve grounding, there is more scope for the participants to reuse lexical forms, as in the following example where the participants do not manage to ground the referent (the move is tagged as `unconfident` and `not_moved`; co-referring expressions are boldfaced):

(4) IF: ***The top of the T*** *faces the right-hand side? Okay?*
IG: (...) ***The top of the T*** *fits next to the first piece, where the the backwards L is.*
IF: ***The top of the T*** *fits next to the first piece?*

As for the IG, low levels of lexical alignment often correspond to cases where an initial referring expression is expanded (e.g., from "*a Z*" to "*a Z but with one end stretched out longer*") or refashioned using a different conceptualisation of the referent (e.g., first trying "*a staircase with a square*" and then "*a zig zag going down on the back*"). These referring strategies, which do not involve high lexical alignment (and often no length decrease either), seem to augment the probability of factual success.

Finally, our predictors selection led us to disregard the direct influence of the *FormAlign* in our model – it is only present within interaction terms for factual success. This suggests a weaker role of the dynamics of syntactic form matching between referring expressions for communicative success, at least when assessed by means of our *FormAlign* measure. In this case, qualitative analysis did not shed light on the interaction of the mean value of *FormAlign* of the IG with its variance.

6 Conclusions

This paper has analysed the dynamics of referring expressions in task-oriented dialogue and investigated their relationship with perceived and factual communicative success. We have introduced three simple measures to quantify length decrease, lexical repetition, and syntactic form matching in coreference chains and applied them to a section of the PentoRef corpus of human-human dialogues annotated with information on referring expressions.

Our descriptive analysis confirms well-known tendencies attested in previous literature: refer-

ring expressions with a common referent tend to decrease and to reuse lexical and syntactic forms more than expected by chance. We have also observed high variability of these results, which we argued is related to a complex interaction of our measures with communicative task success. Although we have operationalised our regression experiment considering success as the dependent variable to be predicted, our study does not assume that alignment *causes* success (in contrast to, for instance, the Interactive Alignment Model proposed by Pickering and Garrod (2004)). Instead, our results hint at more complex relationships between different forms of entrainment and different types of success.

We have shown that lexical alignment has a prevailing role in being indicative of success. However, its relationship to task success is not linear nor symmetric for both participant roles. Qualitative analysis has revealed that there is a connection with the presence or absence of confirmations and clarification requests and with different strategies for proposing referring expressions – very high levels of alignment may be a sign of having reached an impasse in the dialogue, as illustrated by example (4). Achieving a better understanding of what surface forms of referring expressions are related to factual and perceived success, respectively, remains an open issue for future research.

References

Mira Ariel. 1991. The function of accessibility in a theory of grammar. *Journal of Pragmatics*, 16(5):443–463.

Holly P Branigan, Martin J Pickering, and Alexandra A Cleland. 2000. Syntactic co-ordination in dialogue. *Cognition*, 75(2).

Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482.

K. Carbary and M. Tanenhaus. 2011. Conceptual pacts, syntactic priming, and referential form. In *Proceedings of the CogSci Workshop on the Production of Referring Expressions (PRE-CogSci 2011)*.

Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Joachim Daiber, Lautaro Quiroz, Roger Wechsler, and Stella Frank. 2015. Splitting compounds by semantic analogy. In *1st Deep Machine Translation Workshop*, pages 20–28. Charles University in Prague.

Raquel Fernández, David Schlangen, and Tatjana Lucht. 2007. Push-to-talk ain't always bad! comparing different interactivity settings in task-oriented dialogue. In *Proceedings of SemDial 2007*.

Anouschka Foltz, Judith Gaspers, Carolin Meyer, Kristina Thiele, Philipp Cimiano, and Prisca Stenneken. 2015. Temporal effects of alignment in text-based, task-oriented discourse. *Discourse Processes*, 52(8):609–641.

Robert M Krauss and Sidney Weinheimer. 1964. Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1(1-12):113–114.

Robert M. Krauss and Sidney Weinheimer. 1966. Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, 4:343–346.

Kenneth O. McGraw and S. P. Wong. 1992. A common language effect size statistic. *Psychological Bulletin*, 111(2):361–365.

C. Metzing and S. E. Brennan. 2003. When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, 49:237–246.

A. Nenkova, A. Gravano, and J. Hirschberg. 2008. High frequency word entrainment in spoken dialogue. In *Proceedings of ACL*, pages 169–172.

Naho Orita, Eliana Vornov, Naomi Feldman, and Hal Daumé III. 2015. Why discourse affects speakers' choice of referring expressions. In *Proceedings of ACL-IJCNLP*, pages 1639–1649.

M. J. Pickering and S. Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02):169–190.

David Reitter and Johanna D Moore. 2007. Predicting success in dialogue. In *Proceedings of ACL*, pages 808–815.

David Reitter and Johanna D Moore. 2014. Alignment and task success in spoken dialogue. *Journal of Memory and Language*, 76:29–46.

David Schlangen and Raquel Fernández. 2007. Speaking through a noisy channel - experiments on inducing clarification behaviour in human-human dialogue. In *Proceedings of Interspeech*.

David Schlangen and Raquel Fernández. 2008. The Potsdam Dialogue Corpora: Transcription and Annotation Manual. Technical report, University of Potsdam.

Sina Zarrieß, Julian Hough, Casey Kennington, Ramesh Manuvinakurike, David DeVault, Raquel Fernández, and David Schlangen. 2016. PentoRef: A Corpus of Spoken References in Task-oriented Dialogues. In *Proceedings of LREC*.