Компьютерная лингвистика и интеллектуальные технологии:
по материалам международной конференции «Диалог 2016»

Москва, 1–4 июня 2016

ВРЕМЕННАЯ КООРДИНАЦИЯ МЕЖДУ ЖЕСТОВЫМИ И РЕЧЕВЫМИ ЕДИНИЦАМИ В МУЛЬТИМОДАЛЬНОЙ КОММУНИКАЦИИ[1]

Федорова О. В. (olga.fedorova@msu.ru)[1,2,3]
Кибрик А. А. (aakibrik@gmail.com)[1,2]
Коротаев Н. А. (n_korotaev@hotmail.com)[2,3,4]
Литвиненко А. О. (allal1978@gmail.com)[2]
Николаева Ю. В. (julianikk@gmail.com)[1,2]

[1]МГУ имени М. В. Ломоносова, Москва, Россия
[2]Институт языкознания РАН, Москва, Россия
[3]РАНХиГС, Москва, Россия
[4]РГГУ, Москва, Россия

Доклад развивает проблематику мультимодальной лингвистики, рассматривающей все каналы передачи информации — вербальные единицы, просодию, жесты, мимику, направление взора и т. д. — в их взаимодействии. Одним из ключевых вопросов мультимодальных исследований является вопрос о временно́й координации между иллюстративными мануальными жестами (спонтанными жестами рук, сопровождающими речь) и элементарными дискурсивными единицами (базовыми единицами локальной структуры дискурса). В настоящей работе этот вопрос рассматривается на материале создаваемого мультимодального корпуса «Рассказы и разговоры о грушах». В ряде исследований было показано, что время начала жеста обычно опережает время начала речи. Для проверки этого положения был разработан аналитический аппарат, который позволил провести более детальное исследование. В результате выяснилось, что в исследованном материале жесты опережают речь менее чем в половине случаев. Наиболее правдоподобное объяснение полученных отличий от работ предшественников связано с функциональными классами жестов, поскольку распределение жестов по классам существенно зависит от жанра дискурса и индивидуальных особенностей говорящих.

Ключевые слова: мультимодальный дискурс, жест, элементарная дискурсивная единица, временная координация

[1] Работа выполнена при финансовой поддержке РНФ (проект № 14-18-03819).

TEMPORAL COORDINATION BETWEEN GESTURAL AND SPEECH UNITS IN MULTIMODAL COMMUNICATION

Fedorova O. V. (olga.fedorova@msu.ru)[1,2,3]
Kibrik A. A. (aakibrik@gmail.com)[1,2]
Korotaev N. A. (n_korotaev@hotmail.com)[2,3,4]
Litvinenko A. O. (allal1978@gmail.com)[2]
Nikolaeva Ju. V. (julianikk@gmail.com)[1,2]

[1]Lomonosov Moscow State University, Moscow, Russia
[2]Institute of Linguistics RAS, Moscow, Russia
[3]RANEPA, Moscow, Russia
[4]RSUH, Moscow, Russia

This study contributes to the research field of multimodal linguistics. Multimodal linguistics explores numerous channels involved in natural communication, such as verbal structure, prosody, gesticulation, mimics, eye gaze, etc., and treats them as parts of an integral process. Among the key issues in multimodal studies is the question of temporal coordination between the illustrative manual gestures (that is, spontaneous co-speech gestures) and elementary discourse units (that is, basic quanta of the local structure of spoken discourse). We address this issue with the help of a novel multimodal corpus "Pear chats and stories" that is currently under construction. It had been shown in a number of studies that gesture onset usually precedes speech onset. In order to verify this claim through our materials, we developed an analytic method that allowed to conduct a more detailed study. According to our results, it is only less than a half of all gestures that are produced before the corresponding fragment of talk. The most likely explanation of the obtained results is associated with gestures' affiliation in a certain functional class, that is strongly dependent on discourse genre and speakers' individual differences.

Key words: multimodal discourse, gesture, elementary discourse unit, temporal coordination

1. Мультимодальная коммуникация и мультимодальные корпуса

В доминирующей лингвистической традиции общепринято представление, согласно которому основным или даже единственным компонентом языкового взаимодействия является вербальная структура, а другие типы сигнала — в частности, просодия и жесты, — играют второстепенную роль и не относятся к собственно языку. В последние годы, однако, эта точка зрения постепенно

уступает место новой мультимодальной[2] перспективе (Scollon 2006; Кибрик 2010; Knight 2011; Abuczki, Esfandiari 2013; Müller et al. eds. 2014), согласно которой для успешной языковой коммуникации важны и, следовательно, заслуживают внимания все каналы передачи информации: вербальные единицы, просодия, жесты, мимика, направление взора и т. д. В нашем мультимодальном подходе выделяются два вокальных (слуховых) канала — вербальный и просодический, а также группа кинетических (зрительных) каналов (Кибрик 2010). Под вербальным каналом мы понимаем весь речевой материал, который в конечном счете сводится к последовательности фонем. К просодическому каналу относятся несегментные аспекты звука — интонация, дискурсивные акценты, громкость, тембр и т. д. (Кодзасов 2009; Кибрик, Подлесская ред. 2009). К кинетическим каналам (иногда именуемым языком тела) принадлежат мануальные жесты, направление взора, позы и т. д. (Крейдлин 2002; Kendon 2004; McNeill 2005; Николаева 2013).

Мультимодальный корпус — это «аннотированное собрание скоординированной информации из разных коммуникативных каналов, включая речь, направление взора, мануальные жесты и язык тела, которое обычно создается на материале записей человеческого поведения» (Foster, Oberlander 2007: 307–308). В отличие от моноканальных (письменных, основанных на вербальном канале) и мономодальных (речевых, основанных на вербальном и просодическом каналах, принадлежащих к слуховой модальности) корпусов, уже имеющих свою традицию, параметры, по которым можно классифицировать мультимодальные корпуса, еще только вырабатываются. Ниже мы перечислим четыре из них. Самый большой заявленный *объем* мультимодального корпуса — AMI Meeting Corpus, составляет 100 часов (Carletta 2006), однако бóльшая часть данных представлена в нем в виде неразмеченных видеофайлов. ***Характер общения*** собеседников удобно изображать в виде шкалы от контролируемых экспериментов на левом краю до ничем не ограниченного общения на правом. На самом левом краю находится Czech Audio–Visual Speech corpus (Železny et al. 2006), созданный для тестирования системы распознавания речи и включающий 25 часов записи 65 испытуемых, читающих вслух по 200 предложений. Правее расположен Fruit Carts Corpus (Aist et al. 2012), в котором записано 240 видеороликов продолжительностью 4–8 мин. каждый. Испытуемые выполняли стандартное задание — инструктор давал раскладчику инструкции по раскладыванию карточек с нарисованными на них фруктами. Еще правее находится D64 corpus, собранный для изучения социального общения (Campbell 2009), а также InSight Interaction corpus (Brône, Oben 2015), включающий 15 диалогов по 20 мин. каждый. На самом правом краю находятся корпуса, созданные в традиции анализа бытовых диалогов (Mondada 2014). Кроме объема и характера общения, выделяются также такие параметры, как ***количество собеседников*** (2 vs. 3+) и ***среда общения*** (специально созданные условия для

[2] Термин «мультимодальность» опирается на принятое в психологии и нейрофизиологии понимание модальности как принадлежности сигнала к определенной сенсорной системе человека.

проведения записей vs. неподготовленная среда). Три последних параметра важно оценивать с точки зрения естественности коммуникации. Наиболее естественные данные собираются в ходе бытового общения трех и более собеседников в неподготовленной среде. В описываемом ниже корпусе собраны записи общения четырех участников в специально созданных условиях совместного решения некоторой когнитивной задачи.

2. Корпус «Рассказы и разговоры о грушах»

Описываемый корпус является частью более обширного корпуса, создаваемого в рамках проекта РНФ «Язык как он есть: русский мультимодальный дискурс». Корпус включает 24 записи общей продолжительностью около 10 часов и объемом около 110 тыс. словоупотреблений. В качестве стимульного материала при сборе корпуса был использован известный шестиминутный «Фильм о грушах» У. Чейфа (Chafe ed. 1980). Была разработана новая методика сбора материала. В каждой записи принимали участие четыре человека с заранее распределенными ролями. Три коммуниканта — Рассказчик, Комментатор и Пересказчик — участвовали в основной части записи, а четвертый — Слушатель — присоединялся в конце. Сначала Рассказчик и Комментатор смотрели каждый на своем ноутбуке фильм и старались как можно лучше запомнить сюжет и всевозможные детали фильма. Затем Рассказчик излагал содержание фильма Пересказчику, который фильма не видел. На следующем этапе Комментатор дополнял рассказ Рассказчика подробностями, о которых тот не сообщил, при необходимости исправлял его, а Пересказчик уточнял у двух других участников необходимые для последующего пересказа детали; это был этап обсуждения. Наконец, Пересказчик пересказывал содержание фильма Слушателю, который непосредственно перед этим входил в помещение, — это был второй пересказ. После этого Слушатель письменно фиксировал на бумаге услышанный пересказ. Таким образом, основная задача каждого участника заключалась в том, чтобы максимально подробно и понятно донести до других коммуникантов полученную информацию.

При записи речи использовался рекордер ZOOM H6 Handy Recorder с параметрами 96 kHz / 24 bit; речь каждого из трех говорящих записывалась на петличный микрофон SONY ECM–88B; кроме того, отдельно велась общая стереозапись с микрофона рекордера. Три промышленные видеокамеры JAI GO–5000M–USB с частотой 100 к/с и разрешением 1392x1000 записывали крупным планом каждого говорящего в формате mjpeg, который выгодно отличается от других отсутствием межкадрового сжатия. Видеокамера GoPro Hero 4 Black Edition (50 к/с и 2700×1500) записывала общий план. Для регистрации движений глаз были использованы две пары очков–айтрекеров Tobii Glasses II Eye Tracker с частотой 50 Hz и разрешением видеокамеры 1920×1080. Один из айтрекеров был надет на Рассказчика, второй на Пересказчика. Насколько нам известно, подобные айтрекеры еще не использовались при исследовании мультимодального дискурса.

3. Жест и ЭДЕ: опережение, синхронизация или отставание?

3.1. Общие положения

Настоящая работа посвящена одному из ключевых вопросов мультимодальных исследований — ***временно́й координации*** между иллюстративными мануальными жестами и элементарными дискурсивными единицами (ЭДЕ, EDU)[3]. Жесты рук, сопровождающие речь, носят спонтанный характер и не имеют закрепленной формы и фиксированной связи между означаемым и означающим (Николаева и др. 2015). ЭДЕ является базовой единицей локальной структуры дискурса, выделяется на основании преимущественно просодических критериев и прототипически соответствует одной клаузе (Кибрик, Подлесская ред. 2009; Kibrik 2015).

Вопрос о временно́й координации «жест — ЭДЕ» восходит к более масштабному вопросу о том, насколько жесты и речь связаны между собой в когнитивной системе человека. МакНилл утверждает, что жесты и речь одновременно активируются в одном общем источнике, и, следовательно, должны быть синхронизированы как на фонологическом уровне, так и на уровне семантики (выражают один концепт) и прагматики (выполняют одну прагматическую функцию) (McNeill 1992); на эти положения опирается популярная модель Sketch (de Ruiter 2000). С другой стороны, в модели Interface (Kita, Ozyurek 2003) жесты и речь планируются в разных модулях и, соответственно, никаких гипотез о жесторечевой синхронизации в рамках данной модели не выдвигается. Проведенные исследования пока не дают однозначного ответа, какой подход больше соответствует действительности. Так, в работе Ozyurek et al. 2007 авторы, использовав метод вызванных потенциалов мозга, показали, что правильнее говорить не о временно́й, а только о семантической синхронизации, в то время как в работе Loehr 2012 автор говорит о наличии прагматической, структурной и временно́й синхронизации.

Несмотря на отсутствие согласия в вопросе, существует ли общий источник жестикуляции и речи, исследователи сходятся в том, что обычно ***время начала жеста опережает время начала речи***. Эта гипотеза была выдвинута МакНиллом (McNeill 1992), а в последние годы подтверждена на материале английского (Loehr 2012), французского (Ferré 2010) и польского (Karpiński et al. 2009) языков. Одно из возможных объяснений этого феномена состоит в том, что общий когнитивный источник, находящийся на досемантическом уровне, одновременно запускает активацию как абстрактных семантических репрезентаций, так и более конкретных моторных. Однако время, которое тратится на поиск моторных репрезентаций, обычно оказывается меньше, чем время, необходимое для поиска семантических репрезентаций. Это объяснение

---

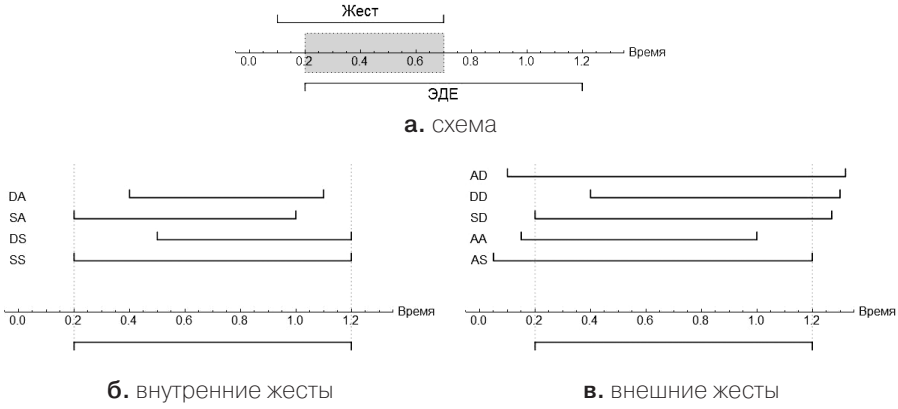
[3] При другом подходе к изучению временно́й координации исследуются отношения между ударными фазами жестов (stroke) и отдельными словами (Schegloff 1984; Leonard, Cummins 2009).

подтверждается в работе Morrel-Samuels, Krauss 1992: чем лучше известно говорящему некоторое слово, тем меньше временно́й интервал между началом жеста и соответствующего фрагмента речи.

Для проверки гипотезы об опережении жеста относительно ЭДЕ[4] мы использовали подкорпус, включающий четыре фрагмента по 6 мин. каждый (4 % всего корпуса); фрагменты были взяты из стадии обсуждения, так что в каждом была представлена речь всех трех собеседников. Аннотации речи и жестов были произведены независимо друг от друга. Для каждой из 1673 выделенных ЭДЕ в программе PRAAT (http://fon.hum.uva.nl/praat/) и для каждого из 614 жестов в программе ELAN (https://tla.mpi.nl/tools/tla-tools/elan/) с точностью до сотой доли секунды были определены время начала и время конца. Для определения временной координации была разработана специальная методика, реализованная в программе Wolfram Mathematica.

3.2. Термины и обозначения

Каждому жесту была поставлена в соответствие определенная ЭДЕ: жест считался соответствующим той ЭДЕ, с которой он имел наибольшее пересечение и которой соответствовал семантически. Длина жеста далее обозначена как L_g, длина ЭДЕ как L_{edu}; их общая часть, обозначенная заливкой на рис. 1, как $L_{g \cap edu}$.



а. схема

б. внутренние жесты

в. внешние жесты

Рис. 1. Временна́я координация «жест–ЭДЕ»

[4] При решении этой задачи необходимо определить, какая из двух единиц — жест или ЭДЕ — первична с точки зрения построения мультимодального дискурса. В работе Kibrik et al. 2015 показано, что ЭДЕ, определяемая на основе общих поведенческих критериев, является центральной единицей мультимодального дискурса. Таким образом, ниже мы рассматриваем вопрос о временно́й координации жеста *по отношению* к ЭДЕ.

Временно́й интервал между началом жеста и началом ЭДЕ мы назвали левым зазором (LGap), временно́й интервал между концом жеста и концом ЭДЕ — правым зазором (RGap). Если жест начинается не раньше ЭДЕ и заканчивается не позже, мы говорим о внутренних жестах (см. рис. 1а) и о внутренних левых (типы DS и DA[5]) и правых (SA и DA) зазорах. Если жест начинается раньше и/или заканчивается позже ЭДЕ, мы говорим о внешних жестах (рис. 1б) и о внешних левых (AS, AA и AD) и правых (SD, DD и AD) зазорах. Таким образом, мы выделяем девять логически возможных типов временно́й координации «жест–ЭДЕ».

При делении пар «жест — ЭДЕ» на типы важен вопрос о точности измерений. Что значит, что границы жеста совпадают с границами ЭДЕ (типы SS, DS, SA, AS и SD на рис. 1)? Мы производили данные вычисления с допустимой погрешностью (т. е. величиной зазора, признаваемого незначимым, далее Δ) в 200 мс, 100 мс и 50 мс, см. рис. 3 ниже.

Временна́я координация «жест — ЭДЕ» была количественно оценена при помощи следующих мер. Точностью (precision, P) мы называем отношение общей части к длине жеста: $L_{g \cap edu} / L_g$. Данная мера показывает, насколько точно жест вписывается в границы ЭДЕ. Полнота (recall, R) — это отношение общей части к длине ЭДЕ. Полнота показывает, насколько жест заполняет ЭДЕ: $L_{g \cap edu} / L_{edu}$. Среднее гармоническое (harmonic mean, HM) рассчитывается по формуле HM = 2PR / (P+R). Эту величину можно сравнить с F_1-мерой, которая используется в алгоритмах извлечения информации.

При исследовании вопроса о временно́й координации в целом мы используем все девять выделенных типов и все три меры, однако в данной работе для проверки гипотезы об опережении жеста относительно ЭДЕ в первую очередь мы учитываем деление жестов на внешние и внутренние, а также результаты измерения точности. Мы различаем ***относительную*** точность P, вычисляемую по вышеприведенной формуле, и ***абсолютную*** точность при измерении Δ в 200 мс, 100 мс и 50 мс. Очевидно, что при определении координации «жест–ЭДЕ» относительная и абсолютная точность могут не совпадать. Например, если $L_{g \cap edu}$ мало, то значение P также будет мало, но если при этом L_g меньше заданной Δ, то получается, что границы этого жеста и ЭДЕ совпадают. Следовательно, для большей надежности результатов нам необходимо использовать оба измерения.

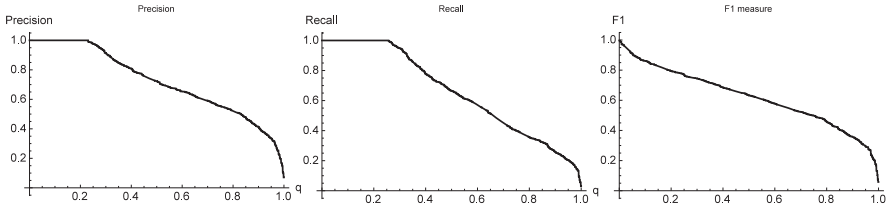
3.3. Результаты

Для каждой пары «жест — ЭДЕ» были вычислены L-Gap, R-Gap, границы $L_{g \cap edu}$, P, R и HM. Сводные данные по P, R и HM представлены на рис. 2.

По графикам видно, что ≈25 % жестов точно входят в границы ЭДЕ, а ≈85 % жестов входят в границы ЭДЕ с P ≥ 0.5, т. е. не менее половины жеста попадает внутрь ЭДЕ (рис. 2а); ≈25 % жестов полностью покрывают ЭДЕ, а 65 % жестов имеют R = 0.5, т. е. заполняют более половины ЭДЕ (рис. 2б); ≈20 % жестов имеют HM = 0.8, а 80 % жестов имеют HM = 0.5 (рис. 2в).

---

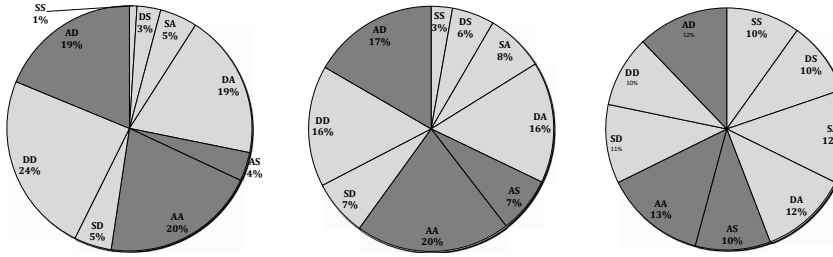
[5] от англ. Anticipation, Synchronization и Delay.



а. точность **б.** полнота **в.** гармоническое среднее

Рис. 2. Меры временно́й координации (по оси x отложены квантили, по оси y — значения соответствующей меры в диапазоне от 0 до 1)

Рассмотрим вопрос о временной координации с другой стороны, распределив пары «жест — ЭДЕ» по девяти типам, приведенным на рис. 1, с Δ в 200 мс, 100 мс и 50 мс, см. рис. 3.



а. Δ = 200 мс **б.** Δ = 100 мс **в.** Δ = 50 мс

Рис. 3. Распределение пар «жест — ЭДЕ» по типам

При Δ в 200 мс (рис. 3а) все типы представлены примерно в равных долях, а число внутренних жестов (SS, DS, SA и DA, 44%) не намного меньше числа внешних. С уменьшением Δ до 100 мс (рис. 3б) и особенно до 50 мс (рис. 3в) распределение предсказуемым образом сдвигается в сторону преобладания тех типов, для которых не требуется строгое соответствие границ жеста и ЭДЕ (то есть типов с наличием одновременно LGap и RGap: DA, AA, DD и AD), а количество внутренних жестов уменьшается до 28%, что уже сравнимо с P, равной 25% (рис 1а). Таким образом, как с абсолютной, так и с относительной точностью только ≈¼ жестов вписывается в границы ЭДЕ. Можно предположить, что остальные ¾ жестов опережают ЭДЕ.

Чтобы проверить гипотезу об опережении жеста относительно ЭДЕ, выделенные типы были поделены на две группы:

(а) начало жеста опережает начало ЭДЕ (AS, AA и AD, на рис. 3 выделены темной заливкой);

(б) начало жеста синхронизировано или отстает от начала ЭДЕ (остальные типы).

Мы видим, что при Δ в 200 мс только 35% жестов попадает в группу (a); при Δ в 100 мс и 50 мс размер группы (a) увеличивается до 44% и 43%, соответственно, однако все равно не достигает 50%. Таким образом, полученные результаты не подтверждают гипотезу об опережающем производстве жестов.

4. Обсуждение результатов и перспективы дальнейших исследований

Итак, мы получили результаты, отличные от результатов предшественников. Какими причинами это может быть обусловлено? Ответить на этот вопрос можно, сравнив наш подкорпус с английским (Loehr 2012), французским (Ferré 2010) и польским (Karpiński et al. 2009) аналогами. Подчеркнем, что в каждой из этих работ авторы использовали просодические единицы, близкие к нашим ЭДЕ: Intonation Phrases в Ferré 2010, Intermediate Phrases в Loehr 2012 и Major Intonation Phrases (MajorIP) в Karpiński et al. 2009. Все эти понятия определяются на основе сравнимых просодических критериев и в общих чертах соответствуют друг другу.

Посмотрим сначала, насколько различаются результаты в целом. В нашем подкорпусе мы получили опережение жеста относительно ЭДЕ в 35% (200 мс), 44% (100 мс) и 43% (50 мс) случаев. Во французском корпусе опережают речь 70% жестов, в польском — 69%, для английского корпуса данные не приведены. Очевидно, что различия весьма велики.

Размер нашего корпуса (24 мин., 1673 ЭДЕ и 614 жестов) заметно превосходит аналоги: английский корпус включал 164 с записи, французский — 244 жеста, а польский — 223 жеста и 773 MajorIP. Серьезным преимуществом нашей работы является также возможность покадрово аннотировать видеозаписи, записанные с частотой 100 к/с. Из трех обсуждаемых корпусов английский был записан с частотой 30 к/с, французский — 24 к/с, для польского корпуса данные не приведены. Что касается измерения Δ, то в нашем подкорпусе данные были проанализированы с Δ в 50 мс, 100 мс и 200 мс. К сожалению, три другие работы не содержат эксплицитных указаний на процедуру подсчета Δ, однако в Karpiński et al. 2009 указано, что в 5% всех случаев начало жеста опережало начало речи меньше, чем на 100 мс, а в 40% случаев — меньше, чем на 200 мс. Кроме того, в Loehr 2012 упоминается, что в среднем время начало жеста опережало время начала речи на 100 мс. По-видимому, абсолютная точность во всех исследованиях была одинаковой или по крайней мере сопоставимой. Таким образом, с точки зрения размера корпуса и точности проводимых измерений наши результаты обладают более высокой степенью валидности.

Мы полагаем, что наиболее правдоподобное объяснение полученных различий связано с функциональными классами жестов. Хорошо известно, что иллюстративные жесты сильно различаются между собой по функциям, которые они выполняют в процессе коммуникации. Согласно популярной классификации МакНилла их можно разделить на: (1) указательные жесты, выполняющие референцию к объекту; (2) иконические жесты, изображающие конкретные

объекты или действия; (3) метафорические жесты, представляющие абстрактные понятия; (4) ритмические жесты, выделяющие фрагменты речи (McNeill 1992); ср. также несколько иную классификацию в работе Николаева 2013.

Французский корпус включал только иконические жесты, остальные корпуса включали все типы жестов. Известно, однако, что распределение жестов по функциональным классам в каждом конкретном случае сильно зависит от жанра дискурса и индивидуальных особенностей говорящих. Так, в частности, ритмические жесты в корпусе МакНилла составляли 44.4% всех жестов (McNeill 1992), в корпусе из работы Theune, Brandhorst 2010 — 32.1%, а в корпусе Николаевой (2013) — всего 15%. Из этих цифр следует, что ритмические жесты, обычно короткие и максимально синхронизированные с речью, могут оказывать сильное влияние на результаты временно́й координации в корпусе в целом. Кроме того, ритмические жесты корректнее анализировать, рассматривая временные отношения между ударными фазами жестов и отдельными словами, а не координацию «жест–ЭДЕ». Известно также, что в диалогической речи количество ритмических жестов увеличивается (Bavelas et al. 1992). Таким образом, на наши результаты могла повлиять разница в количестве ритмических жестов, скоординированных с ударными фазами жестов.

Как представляется, задача развития данного исследования в ближайшем будущем состоит в: (1) увеличении размера подкорпуса; (2) временно́й привязке каждого слова; (3) учете внутренних особенностей ЭДЕ (их синтаксической, коммуникативной и интонационной структуры); (4) разделении жестов на функциональные типы; (5) разделении жестов на подготовительную, ударную и ретракционную фазы (Kendon 1980); (6) анализе временно́й координации пар «жест — ЭДЕ» и «ударная фаза–слово» по разным функциональным типам жестов и типам ЭДЕ. Так, по-видимому, при анализе указательных и ритмических жестов правильнее рассматривать пары «ударная фаза–слово», а при анализе иконических и метафорических жестов — пары «жест — ЭДЕ».

Литература

1. *Abuczki A., Esfandiari B. G.* (2013), An overview of multimodal corpora, annotation tools and schemes, Argumentum, 9, pp. 86–98.
2. *Aist G., Campana E., Allen J., Swift M., Tanenhaus M. K.* (2012), Fruit Carts: A Domain and Corpus for Research in Dialogue Systems and Psycholinguistics, Computational Linguistics, 38 (3), pp. 469–478.
3. *Bavelas J. B., Chovil N., Lawrie D., Wade A.* (1992), Interactive gestures, Discourse Processes, 15 (4), pp. 469–489.
4. *Brône G., Oben B.* (2015), InSight Interaction. A multimodal and multifocal dialogue corpus, Language Resources and Evaluation, 49(1), pp. 195–214.
5. *Campbell N.* (2009), Tools and Resources for Visualising Conversational-Speech Interaction, in M. Kipp et al. (eds.) Multimodal Corpora: From Models of Natural Interaction to Systems and Applications, Springer, Heidelberg.

6. *Carletta J.* (2006), Announcing the AMI Meeting Corpus, The ELRA Newsletter, 11(1), January-March, pp. 3–5.
7. *Chafe W.* (ed.) (1980), The pear stories: Cognitive, cultural, and linguistic aspects of narrative production, Norwood, Ablex.
8. *Chafe W.* (1994), Discourse, consciousness, and time. The flow and displacement of conscious experience in speaking and writing, Chicago.
9. *de Ruiter J. P.* (2000), The production of gesture and speech, in D. McNeill (ed.), Language and Gesture. Cambridge University Press, pp. 248–311.
10. *Ferré G.* (2010). Timing Relationships between Speech and Co-Verbal Gestures in Spontaneous French, in Language Resources and Evaluation, Workshop on Multimodal Corpora, May 2010, Malta, pp. 86–91.
11. *Foster M. E., Oberlander J.* (2007), Corpus-based generation of head and eyebrow motion for an embodied conversational agent, Language Resources and Evaluation, 41 (3/4), pp. 305–323.
12. *Karpiński M., Jarmołowicz-Nowikow E., Malisz Z.* (2009), Aspects of gestural and prosodic structure of multimodal utterances in Polish task-oriented dialogues, Speech and Language Technology, 11, pp. 113–122.
13. *Kendon A.* (1980), Gesticulation and speech: Two aspects of the process of utterance, in M. R. Key (ed.), The relationship of verbal and nonverbal communication, pp. 207–227.
14. *Kendon A.* (2004), Gesture. Visible action as utterance. Cambridge.
15. *Kida T., Faraco M.* (2008), Prédication gestuelle, Faits de Langues, 31–32, pp. 217–226.
16. *Kibrik A. A., Podlesskaya V. I.* (eds.) (2009), Corpus of spoken Russian "Night Dream Stories" [Korpus ustnoy russkoy rechi "Rasskazy o snovideniyakh"], Jazyki slavyanskikh kul'tur, Moscow.
17. *Kibrik A. A.* (2010), Multimodal linguistics [Mul'timodal'naya lingvistika], in Yu. I. Aleksandrov, V. D. Solov'yev (eds.), Cognitive studies [Kognitivnyye issledovaniya], Vol. IV, Institute of psychology, Moscow, pp. 134–152.
18. *Kibrik, A. A.* (2015), The problem of non-discreteness and spoken discourse structure, Computational linguistics and intellectual technologies, 14 (21), vol. 1, pp. 225–233.
19. *Kibrik A., Fedorova O., Nikolaeva Ju.* (2015), Multimodal Discourse: In Search of Units, in G. Airenti, B. Bara, G. Sandini (eds.), Proceedings of the EuroAsian-Pacific Joint Conference on Cognitive Science, 4th European Conference on Cognitive Science, 11th International Conference on Cognitive Science, Torino, Italy, September 25–27, 2015, University of Torino, Torino, pp. 662–667.
20. *Kita S., Ozyurek A.* (2003), What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: evidence for an interface representation of spatial thinking and speaking, Journal of Memory and Language, 48 (1), pp. 16–32.
21. *Knight D.* (2011), Multimodality and active listenership: A corpus approach, Bloomsbury, London.
22. *Kreydlin G. E.* (2002), Nonverbal semiotics [Neverbal'naya semiotika], New literary review, Moscow.

23. *Leonard T., Cummins F.* (2009), Temporal alignment of gesture and speech, in E. Jarmołowicz-Nowikow, K. Juszczyk, Z. Malisz, M. Szczyszek (eds.), Proceedings of GESPIN2009: Gesture and Speech in Interaction, Poznan, Poland, pp. 1–6.
24. *Loehr D.* (2012)б Temporal, structural, and pragmatic synchrony between intonation and gesture, Laboratory Phonology, vol. 3 (1), pp. 71–89.
25. *McNeill D.* (1992), Hand and Mind: What Gestures Reveal about Thought, The University of Chicago Press, Chicago.
26. *McNeill D.* (2005), Gesture and thought, Chicago.
27. *Mondada L.* (2014), Bodies in action, Language and Dialogue, 4 (3), pp. 357–403.
28. *Morrel-Samuels P., Krauss R. M.* (1992), Word familiarity predicts temporal asynchrony of hand gestures and speech, Journal of Experimental Psychology: Human Learning and Memory, 18 (3), pp. 615–622.
29. *Müller C., Fricke E., Cienki A., McNeill D.* (eds.) (2014), Body — Language — Communication, Mouton de Gruyter, Berlin.
30. *Nikolaeva Yu. V.* (2013), Gesticulation in Russian discourse [Illustrativnyye zhesty v russkom diskurse]. Diss. cand. philol. science, MSU, Moscow.
31. *Nikolaeva Yu. V., Kibrik A. A., Fedorova O. V.* (2015), Discourse structure: a perspective from multimodal linguistics, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2015” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii], RGGU, Moscow, pp. 487–499.
32. *Ozyurek A., Willems R. M., Kita S., Hagoort P.* (2007), On-line integration of semantic information from speech and gesture: insights from event-related brain potentials, Journal of Cognitive Neuroscience, 19 (4), pp. 605–616.
33. *Schegloff E. A.* (1984), On some gestures’ relation to talk, in J. M. Atkinson, J. Heritage (eds.), Structures of Social Action, Cambridge University Press, Cambridge, 266–298.
34. *Scollon R.* (2006), Multimodality and the language of politics, in K. Brown (ed.) Encyclopedia of language and linguistics, Elsevier, vol. 9, pp. 386–387.
35. *Theune M., Brandhorst C.* (2010), To beat or not to beat: beat gestures in direction giving, in S. Kopp, I. Wachsmuth (eds.), Gesture in Embodied Communication and Human–Computer Interaction, in Lecture Notes in Artificial Intelligence, vol. 5934. Springer, pp. 195–206.
36. *Železný M., Krňoul Z., Císař P., Matoušek J.* (2006), Design, implementation and evaluation of the Czech realistic audio-visual speech synthesis, Signal Processing, 83 (12), pp. 3657–3673.