EVI: Multilingual Spoken Dialogue Tasks and Dataset for Knowledge-Based Enrolment, Verification, and Identification

Anonymous ACL submission

Abstract

Knowledge-based authentication is crucial for task-oriented spoken dialogue systems that offer personalised and privacy-focused services. Such systems should be able to *enrol* (E), *verify* (V), and *identify* (I) new and recurring users based on their personal information, e.g. postcode, name, and date-of-birth. In this work, we formalise the three authentication tasks and their evaluation protocols, and we present *EVI*, a challenging spoken multilingual dataset with 5,506 dialogues in English, Polish, and French. Our proposed models set the first competitive benchmarks, explore the challenges of multilingual natural language processing of spoken dialogue, and set directions for future research.

1 Introduction

Computer systems need to be able to identify and verify their users before granting access to personalised services and confidential information (Braz and Robert, 2006; O'Gorman, 2003). In particular, **identification (I)** is the process of specifying the identity of a person, i.e. answer the question: *"who are you?"*. On the other hand, **verification (V)** (aka authentication) is the process of confirming the assertion about a claimed identity, i.e. answer *"are you who you claim you are?"* (Jain et al., 2004). In both processes, the system compares information given by the user with information held by the system; thus they presume **enrolment (E)**, that is, the process of registering the identity information of a new user into the system (Jain et al., 2004).

Task-oriented dialogue systems that offer personalised and privacy-focused services (e.g. set up utilities, track a parcel, or access a bank account) should be able to enrol, identify, and verify new and recurring users, without interrupting their natural conversational interface. Different types of authentication factors may be used (Smith, 2001; O'Gorman, 2003): i) knowledge-based ("*what you know*"), rely on a secret *password* or personal information, e.g. full name, date of birth, mother's maiden name, etc.; ii) possession-based ("*what you have*"), rely on possession of a physical *token*, e.g. a smart card, a metal key, etc.; and iii) inherence-based ("*who you are*"), typically rely on *biometric* properties, e.g., a voiceprint, fingerprint, eye scan, or signature (Variani et al., 2014). Most businesses use knowledge-based authentication in their call centres to identify customers over the phone (Hrabí, 2020; Amein, 2020; Petersen, 2019; Morgen, 2012). As conversational AI is increasingly being used to automate call centres, we seek to enable task-oriented spoken dialogue systems with EVI functionalities.

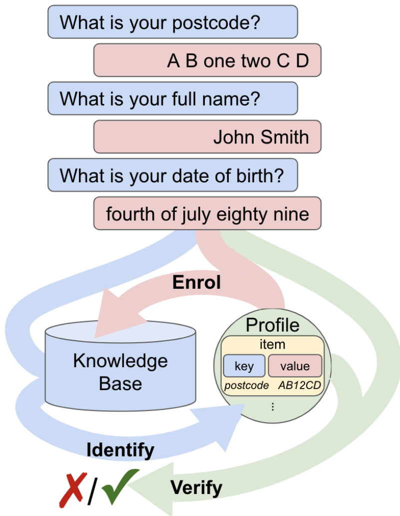

Figure 1: Knowledge-based EVI for task-oriented spoken dialogue systems: enrolment (E) creates a new user profile to store in a KB; identification (I) retrieves a pre-enrolled profile for a user; and verification (V) asserts whether the user matches a claimed profile.

The core **contributions** of this paper are:

1. We motivate and formalise enrolment, verification, and identification as **novel tasks** for task-oriented spoken dialogue systems and propose suitable evaluation protocols (Section 2).
2. We collect and publish a **novel conversational dataset** with 5,506 dialogues that can be used to develop and evaluate EVI-oriented spoken

dialogue systems in 3 languages (British English, Polish, and French; Section 3). The multilingual aspect of the dataset allows us to also study language-specific variations in data and performance, reaching beyond monolingual, English-only setups.

3. We define (Section 4) and evaluate (Section 5) **benchmarks** for the new tasks on the new dataset. Finally, we explore the unique challenges of these tasks and set directions for future research.

The dataset is available online at: [URL].

2 The EVI Dialogue Tasks

Preliminaries. For all tasks, we assume that the dialogue system can interact with a *Knowledge Base (KB)* of stored profiles, $P_{KB} = \{p_1, p_2, ...\}$. Each *profile*, p, is a structured record of a real-world entity (e.g. a user, product, etc.) that comprises one or more *items*, i.e. key-value pairs (e.g. postcode, name, date of birth, etc.). The user and system take alternate turns, t, that make up a multi-turn *dialogue*, $T_{dialogue} = \{t_{1,system}, t_{1,user}, t_{2,system}, t_{2,user}, \dots\}$.

Enrolment Task. The goal of enrolment is to create and store a profile that represents the identity of a *new user* and that can be used to identify or verify the same user in the future. For dialogue-based enrolment, the system must be able to extract all required item key-value pairs from the dialogue to construct a new profile to store in the KB (cf. Fig. 1):

$$p_{new} = \text{enrol}(T_{dialogue}) \tag{1}$$

Verification Task. The goal of verification is to decide whether a user who claims an identity is *genuine* or an *impostor*. For dialogue-based, knowledge-based verification, the system must be able to compare information stored in the KB about the claimed identity with information provided by the user in the dialogue to produce a verification *score* that quantifies the degree of the match (cf. Fig. 1):

$$s_{profile} = \text{verify}(p_{claimed}, T_{dialogue}) \in [0,1], \tag{2}$$

where $s = 1$ signifies a genuine verification attempt, and $s = 0$ denotes an impostor verification attempt. The system designer can apply a threshold, θ, to obtain a crisp verification outcome and control the system's trade-off between security and usability (see later Subsections 4.3 and 4.5).

Identification Task. The goal of identification is to determine the identity of an unknown user from a KB of pre-enrolled user profiles. For dialogue-based, knowledge-base identification, the system must be able to query the KB with the information provided by the user in the dialogue to retrieve a ranked list of the best matching profiles (cf. Fig. 1):

$$p_1, p_2, \ldots = \text{identify}(P_{KB}, T_{dialogue}) \tag{3}$$

The list might be empty if no qualifying profiles (i.e. above a score threshold) could be retrieved.

3 A Multilingual Spoken Dialogue Dataset

We set out to build a novel, first of its kind, human-to-machine conversational dataset that can be used to develop and evaluate task-oriented spoken dialogue systems for all EVI tasks. The dataset is multilingual and covers 3 locales: British English (en-GB), French (fr-FR), and Polish (pl-PL).[1]

3.1 Generating the Profiles Knowledge Base

For each locale, we populate a KB to be shared across EVI tasks. We randomly generated locale-dependent profiles using the *faker* tool.[2] Each profile in the KB consists of its generated item key-value pairs for *postcode*, *full name*, and *date of birth* (cf. Fig. 1). These three different slots are popular in industrial authentication procedures. Table 1 shows the size of the generated KB.

3.2 Collecting the Dialogue Data

We developed a **spoken dialogue system** to collect the postcode, full name, and date of birth of a user over the phone. The system operates under a deterministic policy with static retries for each collection step. We use the same sequence of dialogue acts for all EVI tasks, and vary the scripted prompts (see Subsection 3.3) to elicit more diverse responses:

Q1: What is your postcode?
Q2: Please tell me your postcode.
Q3: I heard [A B 1]. Please tell me your postcode.
Q4: What is your full name?
Q5: Please tell me your first and last name.
Q6: Please spell your full name.
Q7: What is your date of birth?
Q8: Please tell me your date of birth.
Q9: I heard [the 1st of January]. Please tell me your date of birth.

[1] The choice of these languages was motivated by the popularity, the phonetic richness and a large enough base of high-quality crowdworkers.

[2] https://faker.readthedocs.io/; it is a python package that can generate fake but reasonable data (names, addresses, phone numbers, etc.) for bootstrapping databases.

For other locales, see Appendix A. For each locale, we enlisted cohorts of speakers on the *Prolific Academic* (www.prolific.co) crowdsourcing platform. We displayed a random profile from the KB for each speaker to impersonate, e.g.:

Postcode	: *AB1 2CD*	**Kod Pocztowy**	: *12-345*
Full Name	: *John Smith*	**Imię i Nazwisko**	: *Anna Krupa*
Date of Birth	: *4/7/1989*	**Data urodzenia**	: *1/1/2000*

Then, we directed speakers to call a phone number to interact with our spoken dialogue system. To ensure quality, the crowdsourced speakers had to complete all turns of the static policy to receive their payment code.[3] Additionally, we filtered out all dialogues for which text-to-speech detected silence for all turns of a single item or for more than half of the turns of the dialogue.

For each turn, the EVI conversational dataset contains: the unique identifier of the impersonated profile from the KB; a unique speaker identifier; the raw audio data; the n-best list of transcriptions; and any variation in the prompts (see Subsection 3.3). Table 1 shows the size of our dialogue dataset for all locales, which contains 5,506 dialogues in total.

3.3 Speaker Behaviour Analysis

Spoken Dates. To display dates of birth to crowdsourced speakers, we first had to lexicalise them. We used either of two formats at equal proportions:

(*a*) **month=name** : *1 January|stycznia|janvier 2000*
(*b*) **month=number** : *1/1/2000*

These formats acted as *primes* that influenced the speaker's lexical choice. *Priming* is the psychological effect wherein exposure to a stimulus (*prime*) unconsciously influences the response to a later stimulus (*target*). Priming also affects linguistic decision making, e.g. exposure to a lexical item or syntactic structure reinforces reuse of the same pattern in the future (Reitter et al., 2006, 2010). The Sankey diagram[4] in Figure 2 (top) shows that 92% of English speakers primed with the *month=name* format echoed this pattern in Q_7, and only 10% of those switched to say the month's number in follow-up turns (similar results for pl-PL and fr-FR; see Appendix B for their Sankey diagrams).

	counts (unique)	**Locale en-GB**	**pl-PL**	**fr-FR**
KB	#profiles	10,000	10,000	10,000
	#postcodes	2,000	2,000	2,000
	#names(first)	364	153	216
	#names(last)	500	3,455	400
	#names(full)	9,412	9,923	9,433
	#DoBs	8,884	8,862	8,862
Dialogues	#dialogues	1,407	1,991	2,108
	#turns	12,663	17,919	18,972
	#speakers	1,081	803	521
	#profiles	886	961	1,464

Table 1: Size of the EVI Knowledge Bases and Conversational Dataset.

On the other hand, only 54% of English speakers (cf. 26% for pl-PL, 36% for fr-FR; Appendix B) primed with the *month=number* format echoed that pattern in Q_7, and 77% of those switched to say the month's name later. Overall, the *month=name* format (more lexical) had a stronger priming effect than the *month=number* format (more symbolic), and speakers say the month's name (more verbose) increasingly after reprompts (Q_8 and Q_9).

Spoken Spelling. To read back partial spellings of postcodes in the Q_3 reprompts to the speakers, we used either of two strategies at equal proportion:

(*a*) **spell=naive** : *A B one two C D*
(*b*) **spell=nato** :[5] *Alfa Bravo one two Charlie Delta*

These strategies acted as primes that *entrained* the speaker concerning their spelling strategy. *Entrainment* is the phenomenon wherein conversational interlocutors adopt each other's linguistic patterns. Entrainment can be observed at multiple levels, e.g. lexical (Brennan and Clark, 1996), syntactic (Reitter and Moore, 2007), stylistic (Niederhoffer and Pennebaker, 2002), phonetic (Pardo, 2006), and prosodic (Coulston et al., 2002). The Interactive Alignment Model (Pickering and Garrod, 2004) proposes that conversational interlocutors automatically prime each other at multiple levels, causing their speech to converge.[6]

Figure 2 (bottom) shows that only 1% of en-GB speakers spontaneously used NATO spelling before/without encountering the *spell=nato* strategy in Q_3 Conversely, using the *spell=nato* strategy entrained 52% of speakers to adopt that strategy

[3]The workers were not aware that the system was scripted, yielding the natural behaviour of irritated customers.

[4]Sankey diagrams visualise the flow or route of communication (or other quantity) within a system to help locate the most important contributions to a flow. The width of the links between nodes is proportional to the flow rate between them.

[5]The NATO phonetic alphabet substitutes a word for each letter to be easily understood in voice communications; https://www.nato.int/cps/en/natohq/declassified_136216.htm

[6]Alternatively, Communication Accommodation Theory (Giles et al., 1991) proposes that more strategic decisions drive convergence (or divergence).

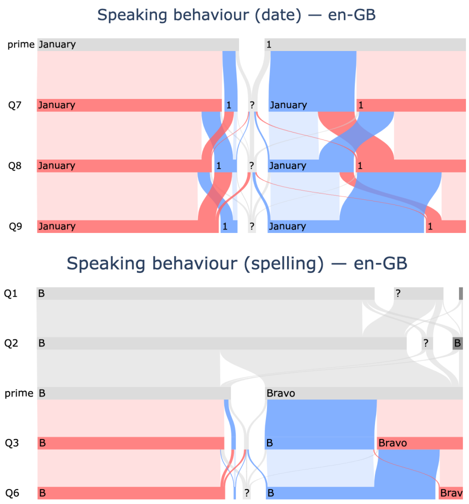

Figure 2: Sankey diagrams that visualise priming and entrainment of speaker behaviour for dates (*top*) and spelling (*bottom*) for the British English locale. Transitions in the direction of priming in red; against, in blue.

in their response to Q_3 Entrainment weakens over time: only 28% of entrained speakers remained entrained by Q_6. Postcodes do not contain letters in the pl-PL and fr-FR locales, so both spelling strategies are equivalent. Only 0.5% of pl-PL and 0.1% of fr-FR speakers spontaneously used complex spelling strategies (listed in Appendix C).

In conclusion, by varying our prompts we increased the variability of speaker behaviours in the dataset. We also corroborate that priming and entrainment are effective tools to subtly guide speaker behaviour towards desired patterns.

4 EVI-oriented Spoken Dialogue Systems

This section presents the components of task-oriented spoken dialogue systems for EVI tasks and provides benchmark implementations for the upcoming experiments (see Sections 5.1, 5.2, and 5.3)

4.1 Components of EVI Dialogue Systems

Automatic Speech Recognition (ASR). When collecting the EVI dataset, we used Google's locale-specific speech-to-text[7] in streaming mode to derive n-best transcriptions and to implement quality control (see Subsection 3.2). Consequently, this is the ASR used in all experiments.

Natural Language Understanding (NLU). For each item, we use an appropriate resource to extract values from the whole ASR n-best list into an NLU results n-best list. In our experiments, we first preprocess to normalise numbers ('one'→'1') and letter spellings ('Bravo|[B for B.*]'→'B'), and then extract values for postcodes using locale-dependent regular expressions ('A(A)9(A) 9AA' for en-GB; '99999' for pl-PL and fr-FR); for names, the lists of names from the US Census[8] and other sources (Remy, 2021); and for dates, the *dateparser* package;[9]. Using these resources, we define two NLU models for value extraction: the `cautious` model requires whole-string match, whereas the `seeking` model searches for (potentially overlapping) substring matches.

Top-Level Policy. All EVI tasks share a common sequence of *dialogue acts* (DAs): the agent asks (request DA) the user to input the value (inform DA) of each profile item successively, with a limited number of re-prompts per item. In the experiments, the order of items is: postcode, full name, and date-of-birth, with up to 3 attempts per item (fixed at the time of dataset collection; see Subsection 3.2).

Task-Level Dialogue Management. Each of the three tasks requires task-specific *dialogue state tracking* (DST) and *dialogue policy*. The DST model tracks and updates the system's state and belief about the values of items and the candidate profiles, whereas dialogue policy selects the following system action (e.g. re-prompt user, proceed to next item, terminate task) and interacts with the profiles KB. We define the task-specific DST models and policies in more detail in Subsections 4.2, 4.3, and 4.4.

Integration with the Profiles KB. For enrolment, the system needs write access to the KB to store the extracted profile; for identification, the system needs read access to the KB to retrieve candidate profiles via a dynamic sequence of queries; and for verification, the claimed profile in the KB is previously made available from an upstream identification process (cf. Fig. 1). In the experiments, we do not explicitly model KB integration for enrolment (write-only access) and verification (downstream of identification); for identification, we model a read-only KB integration that supports querying *by postcode* (exact match) and an `oracle` that always includes the postcode of the correct profile in the query, regardless of the NLU results.

Natural Language Generation (NLG). When collecting the dataset, we used *scripted* prompts (Sub-

[7]https://cloud.google.com/speech-to-text

[8]https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html

[9]https://dateparser.readthedocs.io/ it is a python package that can parse localised dates in any string format

section 3.2) translated for each locale (Appendix A).

Text-to-Speech (TTS). We used Google's[10] locale-specific TTS when collecting the EVI dataset.

4.2 Enrolment Models and Policies

Enrolment DST and Model. We track the value of each item, which is initially *undefined*. After each user input for an item, we may use the NLU n-best results to update its value. When the enrolment policy terminates, the enrolment model straightforwardly builds the new profile from the tracked items. In the experiments, we update an item's value with its latest *top-1* result of the NLU (if not empty).

Enrolment Policy. The task-level policy determines when to proceed to the next item, and decides when to terminate enrolment. The policy (re)prompts the user about an item until either the DST returns a well-defined value or the top-level policy reaches the limit for attempts (3; see Subsection 4.1). After exhausting all items, the policy terminates and writes the new profile into the KB.

4.3 Verification Models and Policies

Verification DST and Model. We track a verification score for each item s_{item} as follows (cf. Eq. 2):

$$s_{item} = \textbf{score}(\text{item}(p_{claimed}), \text{item}(T_{dialogue})) \in [0,1], \quad (4)$$

The scores are initially undefined, and we track their maximum evaluation after each user input. For the experiments, we define the following scoring models: the `random` model samples from the $[0,1]$ uniform distribution; the `exact` model returns 1 if the value from the claimed profile exactly matches any NLU n-best result, else, 0 (*undefined* for no NLU results); and the `fuzzy` model returns the best *fuzzy match* score between the value from the claimed profile and all NLU n-best results (*undefined* for no NLU results). We implement this as the normalised Levenshtein edit distance using the Wagner–Fischer algorithm (Wagner and Fischer, 1974). Finally, we evaluate a logical expression under *fuzzy logic* to combine all item-level scores (Eq. 4) into a profile-level score as follows (see Eq. 2):

$$\begin{aligned} s_{profile} = s_{postcode} \text{ AND } s_{dob} \text{ AND} \\ (s_{name\_full} \text{ OR}(s_{name\_first} \text{ AND } s_{name\_last})) \end{aligned} \quad (5)$$

Fuzzy logic (Zadeh, 1996) is a many-valued logic wherein truth values are real numbers in $[0,1]$ that represent degrees of truthfulness and reasons using fuzzy logic operators (analogous to Boolean logic's AND, OR, and NOT). In the experiments, we choose the standard fuzzy logic operators (Zadeh, 1996):

$$\begin{aligned} &\textbf{Boolean} \longleftrightarrow \textbf{Fuzzy} \\ &\text{AND}(x,y) \longleftrightarrow \min(x,y) \\ &\text{OR}(x,y) \longleftrightarrow \max(x,y) \\ &\text{NOT}(x) \longleftrightarrow 1-x \end{aligned} \quad (6)$$

Verification Policy. The task-level policy determines when to proceed to the next item, and decides when to terminate the verification process. The policy (re)prompts the user about an item until either the DST returns a well-defined score (Eq. 4) or the top-level policy reaches the limit for attempts (again, 3). The policy terminates either after exhausting all items or when it meets an *early termination* criterion: a low upper bound on the profile score (i.e. Eq. 5 with *undefined* $\equiv 1$ is below the verification threshold, θ) guarantees a negative verification outcome. Upon termination, the policy returns the profile-level verification score (Eq. 5 with *undefined* $\equiv 0$).

4.4 Identification Models and Policies

Identification DST and Model. We track the NLU n-best results from all turns and the candidate profiles retrieved from the KB. Our identification process is an *anytime* algorithm (Zilberstein, 1996) that ranks the thus-far retrieved profiles by a score (Eq. 5), excluding profiles below an identification threshold, θ. Following the literature on *fuzzy retrieval* (Zadrożny and Nowacka, 2009), instead of the standard fuzzy operators (Eq. 6), we use `p-norm` fuzzy operators (Salton et al., 1983):[11]

$$\begin{aligned} \text{AND}^p(s_1,\ldots,s_n) &= 1 - \left(\frac{1}{n}\sum_{i=1}^{n}|1-s_i|^p\right)^{1/p} \\ \text{OR}^p(s_1,\ldots,s_n) &= \left(\frac{1}{n}\sum_{i=1}^{n}|s_i|^p\right)^{1/p} \end{aligned} \quad (7)$$

In the experiments, we approximate Eq. 7 by the *infinity-one* linear combination (Smith, 1990):

$$\begin{aligned} \text{OR}_\alpha &= \alpha \text{OR}^\infty + (1-\alpha)\text{OR}^1 \\ &= \alpha \max + (1-\alpha)\text{mean} \\ \text{AND}_\alpha &= \alpha \text{AND}^\infty + (1-\alpha)\text{AND}^1 \\ &= \alpha \min + (1-\alpha)\text{mean} \end{aligned} \quad (8)$$

Note that $\text{AND}_1 = \text{AND}^\infty = \min$ and $\text{OR}_1 = \text{OR}^\infty = \max$ are the standard fuzzy operators (Eq. 6). Finally, an identification `oracle` always retrieves the correct profile if it is among the tracked candidates (i.e. retrieved from the KB).

[10] https://cloud.google.com/text-to-speech

[11] The expression is based on the L^p-norm, $||x||_p := \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$, and is related to the generalised (aka power or Hölder) means (Bullen, 2013).

	models	Profile				Postcode				Name				DoB			
	nlu	P%	R%	F1%	L	P%	R%	F1%	L	P%	R%	F1%	L	P%	R%	F1%	L
en-GB	cautious	**38.83**	**30.27**	**34.02**	4.15	**69.08**	**55.20**	**61.37**	1.83	**65.88**	**64.88**	**65.38**	1.12	**80.37**	**78.97**	**79.66**	1.21
	seeking	27.44	23.34	25.22	**3.86**	59.90	51.16	55.18	**1.70**	63.74	63.51	63.63	**1.10**	63.86	63.58	63.72	**1.07**
pl-PL	cautious	**66.41**	**60.37**	**63.25**	3.98	**95.51**	**91.91**	**93.68**	1.51	**71.86**	**69.26**	**70.54**	**1.20**	**92.92**	**90.31**	**91.59**	1.26
	seeking	53.07	51.63	52.34	**3.69**	87.85	86.44	87.14	**1.38**	69.76	69.16	69.46	**1.20**	82.83	82.37	82.60	**1.11**
fr-FR	cautious	**34.22**	**30.37**	**32.19**	3.85	**77.62**	**72.09**	**74.75**	1.50	44.21	44.00	44.10	**1.06**	**90.81**	**86.81**	**88.76**	1.29
	seeking	26.46	24.68	25.54	**3.63**	75.03	70.43	72.66	**1.46**	**44.27**	**44.19**	**44.23**	**1.06**	72.12	71.57	71.84	**1.10**

Table 2: Results for enrolment task: Precision (P), Recall (R), F1 score, and average number of turns (L) for exact match of the whole profile and each of its items (postcode, full name, and date of birth (DoB)).

Identification Policy. The task-level policy queries the KB to retrieve candidate profiles (see Subsection 4.1), determines when to proceed to the next item, and decides when to terminate the identification process. The policy queries the KB with the NLU n-best results, and sends the retrieved profiles to the DST. Similarly to verification, the policy (re)prompts the user about an item until either the DST returns a well-defined score (Eq. 4) or the top-level policy reaches the limit for attempts (again, 3). The policy terminates after having exhausted all items, or when the anytime result of identification is an empty list and the KB cannot be queried by any upcoming item. Upon termination, the policy returns the ranked list of identified profiles.

4.5 Evaluating the EVI Tasks

Evaluating Enrolment. Suitable evaluation metrics come from the area of information extraction: *precision* (P), *recall* (R), and *F1* score, at the profile level or per item.[12]

Evaluating Verification. The relevant literature describes two basic metrics (El-Abed et al., 2012): *False Rejection Rate* (FRR) is the proportion of genuine users that the system incorrectly rejects as impostors; conversely, *False Acceptance Rate* (FAR) is the proportion of impostors that the system incorrectly accepts as genuine. Lower FRR indicates more usable systems, and lower FAR, more secure, e.g. FRR $= 1\%$ at FAR $= 1/10\,000$ means that 1% of genuine users will fail verification at the security level that falsely accepts 1 impostor per $10{,}000$ impostor attempts. *Equal Error Rate* (EER) is the error rate when FAR $=$ FRR; it is a popular evaluation metric when a security level is not a priori specified. Finally, the *Detection Error Trade-off* (DET) graph plots FRR (y-axis) against FAR (x-axis) for varying values of the verification threshold (θ) to visualise usability across a range of security levels (Martin et al., 1997).

[12]Enrolment outputs (new profiles) are stored in the KB and feed into I&V downstream tasks (Fig. 1); evaluating interactions among tasks is outside the scope of this paper.

	Turns (Subsection 3.2)	Postcode P%	Postcode R%	Postcode F1%	Name P%	Name R%	Name F1%	DoB P%	DoB R%	DoB F1%
en-GB	single(Q_i), $i=1,4,7$	68.17	32.80	44.29	**67.35**	61.71	64.40	81.48	69.00	74.73
	single(Q_i), $i=2,5,8$	73.27	39.02	50.92	65.47	56.72	60.78	79.64	66.98	72.76
	single(Q_i), $i=3,6,9$	**75.95**	37.64	50.34	20.03	10.26	13.57	**86.31**	71.97	78.49
	multi (Q_{1-9})	69.08	**55.20**	**61.37**	65.88	**64.88**	**65.38**	80.37	**78.97**	**79.66**
pl-PL	single(Q_i), $i=1,4,7$	95.95	58.26	72.50	**74.11**	62.98	68.10	93.69	76.04	83.95
	single(Q_i), $i=2,5,8$	97.37	79.96	87.81	73.62	62.08	67.36	93.33	77.30	84.56
	single(Q_i), $i=3,6,9$	**97.53**	85.33	91.03	21.95	6.68	10.24	**93.80**	81.27	87.08
	multi (Q_{1-9})	95.51	**91.91**	**93.68**	71.86	**69.26**	**70.54**	92.92	**90.31**	**91.59**
fr-FR	single(Q_i), $i=1,4,7$	80.76	51.59	62.96	**45.06**	42.86	43.93	91.21	73.42	81.36
	single(Q_i), $i=2,5,8$	82.48	65.02	72.72	41.44	39.72	40.56	**92.91**	74.61	82.76
	single(Q_i), $i=3,6,9$	**83.09**	65.07	72.98	2.64	1.85	2.18	92.02	76.08	83.29
	multi (Q_{1-9})	77.62	**72.09**	**74.75**	44.21	**44.00**	**44.10**	90.81	**86.81**	**88.76**

Table 3: Results for single- vs multi-turn value extraction with `cautious` NLU: Precision (P), Recall (R), F1 score per item (postcode, full name, and date of birth).

Evaluating Identification. We rely on the *identification rate at rank r* (IR@r) (El-Abed et al., 2012): the proportion of identification transactions by pre-enrolled users in which the correct profile is among the top-r retrieved by the system. It is equivalent to the familiar *recall at rank* metric from information retrieval (Manning et al., 2008).

5 Experiments and Results

This section evaluates benchmarks and empirically explores the unique challenges of each EVI task.

Experimental Setup. For all experiments, we deterministically simulate ground truths and user inputs from our EVI KB and dataset, respectively (see Subections 3.1 and 3.2). The implementations of ASR, top-level policy, NLG, and TTS were set at the time of data collection and are common for all EVI tasks (see Subsection 4.1). Subsection 4.5 describes the evaluation metrics for each task.

5.1 Enrolment Experiments

We evaluate the enrolment policy with `cautious` or `seeking` NLU (see Subsection 4.1).

Results. Table 2 shows the impact of NLU on enrolment task accuracy (i.e. precision, recall, F1), for the whole profile and per item, and the average dialogue length. For whole profiles and almost all items, `cautious` NLU, which is more conservative and extracts fewer values, yields better accuracy than `seeking` NLU, which is more liberal and over-extracts values. Notably, extraction of French names

models		en-GB			pl-PL			fr-FR		
nlu	V-model	EER%	FRR%	L	EER%	FRR%	L	EER%	FRR%	L
cautious	random	32.95	54.70	4.15 (2.85)	17.28	30.99	3.98 (2.67)	22.50	49.83	3.85 (2.38)
cautious	exact	28.22	56.42	4.15(2.78)	17.60	35.20	3.98 (2.59)	27.48	54.95	3.85 (2.30)
cautious	fuzzy	22.47	24.27	4.15 (3.09)	6.88	11.24	3.98 (2.76)	11.01	29.06	3.85 (2.57)
seeking	random	31.86	58.67	3.86 (2.59)	17.83	38.93	3.69 (2.37)	24.11	49.22	3.63 (2.30)
seeking	exact	30.89	61.77	3.86 (2.50)	21.15	42.29	3.69 (2.31)	25.87	51.73	3.63 (2.25)
seeking	fuzzy	**11.27**	**21.06**	3.86 (2.84)	**4.27**	**10.56**	3.69 (2.53)	**9.11**	**18.73**	3.63 (2.53)

Table 4: Results of verification task: Equal Error Rate (EER), False Rejection Rate (FRR) @FAR = 1/10,000, and average number of turns (L; in parentheses: with early termination @FAR = 1/10,000).

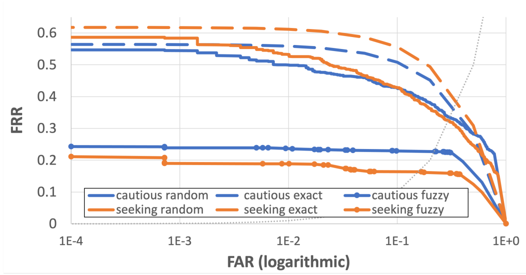

Figure 3: Detection Error Trade-off (DET) curves for the en-GB locale. A curve that is closer to the bottom of the plot corresponds to better verification performance.

and English postcodes (alphanumeric) was less accurate than for other locales (digit-only postcodes).

Further Analysis. Table 3 shows the pre-item accuracy (i.e. precision, recall, F1) of single- and multi-turn value extraction with the `cautious` model. Consistently, recall with multi-turn extraction is higher than single-turn recall of any individual turn. Conversely, individual single-turns yield the highest precisions. Across locales, the relevant precisions of turns is retained for postcodes ($Q_3 > Q_2 > Q_1$) and names ($Q_4 > Q_5 > Q_6$) (cf. Section 3.2). In particular, extraction of name spellings (Q_6) is distinctly poor; this barely affects multi-turn performance, because, on average, the system collects names before Q_6 (Table 2).

5.2 Verification Experiments

We evaluate the verification policy with `cautious` or `seeking` NLU and `random`, `exact`, or `fuzzy` verification (Subsection 4.3) on the EVI dataset and KB (Section 3), from which we sample genuine and impostor profiles at a 1 : 1 ratio.

Results. Table 4 shows the impact of NLU and verification models on the equal error rate (EER), the FRR at the FAR $= 1/10\ 000$ security level and length. Consistently, `seeking` NLU with `fuzzy` verification yields the best EER and FRR. Interestingly, `exact` verification fails to improve reliably over the `random` baseline. Finally, early termination shortens verification length by 25-30%.

Further Analysis. Figure 3 shows the DET curves for the en-GB locale and all models. `Exact` verification produces single points on the y-axis, which we linearly interpolate to produce its DET curve. Again, `seeking` NLU with `fuzzy` verification yields the best usability-security trade-off (lowest-lying curve) for the whole range of security levels in the graph. The same holds for the DET curves of the pl-PL and fr-FR (shown in Appendix D).

5.3 Identification Experiments

We evaluate the identification policy with `cautious` or `seeking` NLU (Subsection 4.1), and no (`none`), `exact`, `fuzzy`, or `oracle` (upper bound) identification (Subsection 4.4). We vary the α parameter of the infinity-one `p-norm` (Eq. 7).

Results. Table 5 shows the impact of NLU and identification models on identification rate at rank 1 and identification length. Without an explicit identification model (`none`) the agent cannot differentiate among multiple retrieved profiles and accuracy is very low. Consistently, `seeking` NLU, `fuzzy` models, and $\alpha = 0.5$ perform better than `cautious` NLU, `exact` matching, and $\alpha = 1$ (i.e. the standard fuzzy operators), respectively. These effects are orthogonal: `seeking` NLU with `fuzzy` model and $\alpha = 0.5$ produces the best accuracy, almost on par with the `oracle`.

Further Analysis. Most identification errors ($> 98\%$) were caused by low recall: the correct target profile was not included in those returned by querying the KB with the NLU results, which is reminiscent of the *unlinkable entity* (NIL) problem from entity linking (Ling et al., 2015; Hoffart et al., 2014; McNamee and Dang, 2009). Table 6 shows the upper bounds using a `KB oracle` (Subsection 4.1), and corroborates the results of Table 5. The best combination (`seeking` NLU, `fuzzy` model and $\alpha = 0.5$) can achieve almost perfect performance as an upper bound.

models		en-GB		pl-PL		fr-FR	
nlu	**I-model**	**IR@1**	**L**	**IR@1**	**L**	**IR@1**	**L**
cautious	none	9.90	3.64	19.74	3.86	14.95	3.62
seeking	none	10.04	3.54	19.89	3.71	15.09	3.46
cautious	exact($\alpha=1$)	50.22	3.64	65.90	3.86	48.50	3.62
cautious	fuzzy($\alpha=1$)	64.88	3.64	89.15	3.86	71.00	3.62
seeking	exact($\alpha=1$)	46.75	3.54	61.93	3.71	52.40	3.46
seeking	fuzzy($\alpha=1$)	66.18	3.54	93.82	3.71	79.73	3.46
cautious	exact($\alpha=0.5$)	66.11	3.64	94.22	3.86	79.31	3.62
cautious	fuzzy($\alpha=0.5$)	66.33	3.64	94.32	3.86	78.97	3.62
seeking	exact($\alpha=0.5$)	67.27	3.54	94.88	3.71	80.35	3.46
seeking	fuzzy($\alpha=0.5$)	**67.77**	3.54	**95.13**	3.71	**80.83**	3.46
cautious	*oracle*	66.55	2.12	94.37	1.56	80.92	1.75
seeking	*oracle*	67.99	2.09	95.38	1.52	81.02	1.73

Table 5: Results of identification task: Identification Rate at rank 1 (IR@1) and average dialogue length (L).

5.4 Directions for Further Research

Our findings highlight the most promising directions for further improvements. In particular, for enrolment: high-precision NLU and multi-turn belief tracking; for verification: high-recall NLU and fuzzy matching; and for identification: high-recall NLU, fuzzy retrieval, and boosting the recall of querying the KB. All tasks can benefit from better multilingual NLU, and our dataset includes audios to encourage improvements in speech-to-text.

6 Related Work

Authentication Tasks. Our EVI tasks seek to automate the process of knowledge-based authentication (Braz and Robert, 2006; O'Gorman, 2003) in a voice communication context (O'Gorman et al., 2006a,b; O'gorman et al., 2005) using task-oriented spoken dialogue systems. We define and evaluate the tasks analogously to automated systems for biometric authentication (signatures, Yeung et al., 2004; fingerprints, Maio et al., 2002; faces, Phillips et al., 2003; irides, Phillips et al., 2008; and voice, Doddington et al., 2000).

Dialogues, NLP, and Logic. Our EVI benchmarks focus on speech recognition and spoken language understanding of names (Kaplan, 2020; Pappu and Rudnicky, 2014), dates (Price et al., 2021), and spellings (Vertanen and Kristensson, 2012; Filisko and Seneff, 2004; Chung et al., 2003). Furthermore, enrolment is a particular case of the slot-filling dialogue task (Young, 2002; Bellegarda, 2014); and identification is related to information retrieval and shares challenges with entity linking (Ling et al., 2015; Hoffart et al., 2014; McNamee and Dang, 2009). We extend fuzzy logic methods from information retrieval (Radecki, 1979; Zadrożny and Nowacka, 2009; Salton et al., 1983) and from multi-modal verification (Lau et al., 2004; Conti et al., 2007; Azzini et al., 2007) to the context of spoken dialogues.

models		en-GB		pl-PL		fr-FR	
nlu	**I-model**	**IR@1**	**L**	**IR@1**	**L**	**IR@1**	**L**
seeking	none	15.53	3.86	20.54	3.69	18.46	3.63
seeking	exact($\alpha=1$)	38.22	3.86	57.71	3.69	48.27	3.63
seeking	fuzzy($\alpha=1$)	81.86	3.86	95.63	3.69	90.18	3.63
seeking	exact($\alpha=0.5$)	96.60	3.86	97.79	3.69	97.63	3.63
seeking	fuzzy($\alpha=0.5$)	**98.19**	3.86	**98.74**	3.69	**98.81**	3.63
seeking	*oracle*	100.00	1.00	100.00	1.00	100.00	1.00

Table 6: Identification task with a `KB oracle`.

Dialogue Datasets. Research in dialogue systems is driven by competitions (Kim et al., 2019; Gunasekara et al., 2020) and challenge datasets, which may be human-to-human (Schrading et al., 2015; Lowe et al., 2015; Ritter et al., 2010), machine-to-machine (Shah et al., 2018), or human-to-machine (H2M) conversations; about single (Coope et al., 2020; Wen et al., 2017; Hemphill et al., 1990) or multiple domains (Rastogi et al., 2020; Zhu et al., 2020; Zang et al., 2020; Budzianowski et al., 2018; El Asri et al., 2017); in one or several languages (Xu et al., 2020; Li et al., 2021); and with written or spoken data (Lugosch et al., 2019; Li et al., 2018; Hemphill et al., 1990). Our EVI dataset is a spoken-language, multi-lingual, single-domain, human-to-machine challenge dataset for multiple tasks, which were not covered by any dialogue dataset from prior work.

7 Conclusion

We introduced novel spoken-dialogue tasks (knowledge-based enrolment, verification, and identification), the EVI multi-lingual spoken-dialogue dataset with 5,506 dialogues, and benchmark models, evaluations, and upper-performance bounds that leave ample margins for future improvements.

Limitations. During data collection, our policy (fixed-length with reprompts for all items) might have caused artefacts in speaker behaviour (e.g. frustration, chuckling, simplification for later items). Additionally, speaker behaviour of crowd-sourced speakers who impersonate a fake profile will be qualitatively different to presenting one's own personal information; however, ethical and privacy concerns preclude the publication of a dataset with real data. Finally, our current evaluation considers each task in isolation, although in practice they form a sequence (enrolment, identification, and then verification) that may propagate errors.

Future Work. We invite the community to work on the novel EVI tasks and challenge dataset, which pose a variety of unresolved technical challenges: speech recognition, multi-turn spoken language understanding, fuzzy matching and retrieval, etc.

Ethical Considerations

[INSTITUTION-ANONYMOUS] is ISO27k-certified and fully GDPR-compliant. Before data collection, we informed the crowd-sourced human workers that their voluntary participation will allow us to collect, store, publish, and use their fully-anonymous data for research purposes. During data collection, we did not ask workers for their own personal information (e.g. name, postcode); instead, we provided fictional (but realistic looking) profiles for them to impersonate. We instructed workers on how to hide their caller id, we did not store any in-bound phone numbers, and we use fully anonymised identifiers in our dataset. Finally, we offered a fair compensation (around the average hourly wage in the US and the UK, pro-rata) to all workers.

References

John Amein. 2020. Hidden risks of consumer-grade biometrics. *Biometric Technology Today*, 2020(10):5–8.

Antonia Azzini, Stefania Marrara, Roberto Sassi, and Fabio Scotti. 2007. A fuzzy approach to multimodal biometric authentication. In *Proceedings of KES*.

Jerome R Bellegarda. 2014. Spoken language understanding for natural interaction: The siri experience. *Natural interaction with robots, knowbots and smartphones*, pages 3–14.

Christina Braz and Jean-Marc Robert. 2006. Security and usability: the case of the user authentication methods. In *Proceedings of l'Interaction Homme-Machine*.

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.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of EMNLP*.

Peter S Bullen. 2013. *Handbook of means and their inequalities*, volume 560. Springer Science & Business Media.

Grace Chung, Stephanie Seneff, and Chao Wang. 2003. Automatic acquisition of names using speak and spell mode in spoken dialogue systems. In *Proceedings of NAACL-HLT*.

Vincenzo Conti, Giovanni Milici, Patrizia Ribino, Filippo Sorbello, and Salvatore Vitabile. 2007. Fuzzy fusion in multimodal biometric systems. In *Proceedings of KES*.

Samuel Coope, Tyler Farghly, Daniela Gerz, Ivan Vulić, and Matthew Henderson. 2020. Span-convert: Few-shot span extraction for dialog with pretrained conversational representations. In *Proceedings of ACL*.

Rachel Coulston, Sharon Oviatt, and Courtney Darves. 2002. Amplitude convergence in children's conversational speech with animated personas. In *Proceedings of ICSLP*.

George R Doddington, Mark A Przybocki, Alvin F Martin, and Douglas A Reynolds. 2000. The nist speaker recognition evaluation–overview, methodology, systems, results, perspective. *Speech communication*, 31(2-3):225–254.

Mohamad El-Abed, Romain Giot, Baptiste Hemery, and Christophe Rosenberger. 2012. Evaluation of biometric systems: A study of users' acceptance and satisfaction. *International Journal of Biometrics*, 4(3):265–290.

Layla El Asri, Hannes Schulz, Shikhar Kr Sarma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of SIGDIAL*.

Edward Filisko and Stephanie Seneff. 2004. Error detection and recovery in spoken dialogue systems. In *Proceedings of HLT-NAACL Workshop on Spoken Language Understanding for Conversational Systems and Higher Level Linguistic Information for Speech Processing*.

Howard Giles, Nikolas Coupland, and IUSTINE Coupland. 1991. 1. accommodation theory: Communication, context, and. *Contexts of accommodation: Developments in applied sociolinguistics*, 1.

Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, et al. 2020. Overview of the ninth dialog system technology challenge: Dstc9. *arXiv preprint arXiv:2011.06486*.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Proceedings of the Workshop on Speech and Natural Language*. HLT '90.

Johannes Hoffart, Yasemin Altun, and Gerhard Weikum. 2014. Discovering emerging entities with ambiguous names. In *Proceedings of TheWebConf*.

Michal Hrabí. 2020. Call centres: going voice-first in the post-covid world. *Biometric Technology Today*, 2020(8):10–12.

Anil K Jain, Arun Ross, and Salil Prabhakar. 2004. An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, 14(1):4–20.

Micaela Kaplan. 2020. May i ask who's calling? named entity recognition on call center transcripts for privacy law compliance. In *Proceedings of the EMNLP Workshop on Noisy User-generated Text (WNUT)*.

Seokhwan Kim, Michel Galley, Chulaka Gunasekara, Sungjin Lee, Adam Atkinson, Baolin Peng, Hannes Schulz, Jianfeng Gao, Jinchao Li, Mahmoud Adada, et al. 2019. The eighth dialog system technology challenge. *arXiv preprint arXiv:1911.06394*.

Chun Wai Lau, Bin Ma, Helen Mei-Ling Meng, Yiu-Sang Moon, and Yeung Yam. 2004. Fuzzy logic decision fusion in a multimodal biometric system. In *Proceedings of ICSLP*.

Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. 2018. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. *arXiv preprint arXiv:1804.00320*.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of EACL*.

Xiao Ling, Sameer Singh, and Daniel S Weld. 2015. Design challenges for entity linking. *TACL*, 3:315–328.

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of SIGDIAL*.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech model pre-training for end-to-end spoken language understanding. *arXiv preprint arXiv:1904.03670*.

Dario Maio, Davide Maltoni, Raffaele Cappelli, James L. Wayman, and Anil K. Jain. 2002. Fvc2000: Fingerprint verification competition. *IEEE transactions on pattern analysis and machine intelligence*, 24(3):402–412.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval, Chapter 8*. Cambridge University Press.

Alvin Martin, George Doddington, Terri Kamm, Mark Ordowski, and Mark Przybocki. 1997. The det curve in assessment of detection task performance. Technical report, NIST.

Paul McNamee and Hoa Trang Dang. 2009. Overview of the tac 2009 knowledge base population track. In *Proceedings of TAC*.

Bob Morgen. 2012. Voice biometrics for customer authentication. *Biometric Technology Today*, 2012(2):8–11.

Kate G Niederhoffer and James W Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

Lawrence O'Gorman. 2003. Comparing passwords, tokens, and biometrics for user authentication. volume 91, pages 2021–2040. IEEE.

Lawrence O'gorman, Amit Bagga, and Jon Bentley. 2005. Query-directed passwords. *Computers & Security*, 24(7):546–560.

L O'Gorman, L Brotman, and M Sammon. 2006a. Comparing authentication protocols for securely accessing systems by voice. In *Proceedings of ICSKM*.

Lawrence O'Gorman, Lynne Brotman, and Michael Sammon. 2006b. How to speak an authentication secret securely from an eavesdropper. In *International Workshop on Security Protocols*. Springer.

Aasish Pappu and Alexander Rudnicky. 2014. Knowledge acquisition strategies for goal-oriented dialog systems. In *Proceedings of SIGDIAL*.

Jennifer S Pardo. 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4):2382–2393.

John Petersen. 2019. The complexity of consent and privacy in biometrics–worldwide. *Biometric Technology Today*, 2019(8):5–7.

P Jonathon Phillips, Kevin W Bowyer, Patrick J Flynn, Xiaomei Liu, and W Todd Scruggs. 2008. The iris challenge evaluation 2005. In *Proceedings of the International Conference on Biometrics*. IEEE.

P Jonathon Phillips, Patrick Grother, Ross Micheals, Duane M Blackburn, Elham Tabassi, and Mike Bone. 2003. Face recognition vendor test 2002. In *Proceedings of the International SOI Conference*. IEEE.

Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

Ryan Price, Mahnoosh Mehrabani, Narendra Gupta, Yeon-Jun Kim, Shahab Jalalvand, Minhua Chen, Yanjie Zhao, and Srinivas Bangalore. 2021. A hybrid approach to scalable and robust spoken language understanding in enterprise virtual agents. In *Proceedings of NAACL-HLT*.

Tadeusz Radecki. 1979. Fuzzy set theoretical approach to document retrieval. *Information Processing & Management*, 15(5):247–259.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of AAAI*.

David Reitter, Frank Keller, and Johanna D Moore. 2006. Computational modelling of structural priming in dialogue. In *Proceedings of NAACL-HLT*.

David Reitter and Johanna D Moore. 2007. Predicting success in dialogue. In *Proceedings of ACL*.

David Reitter, Johanna D Moore, and Frank Keller. 2010. Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In *Proceedings of CogSci*.

Philippe Remy. 2021. Name dataset. https://github.com/philipperemy/name-dataset.

Alan Ritter, Colin Cherry, and William B Dolan. 2010. Unsupervised modeling of twitter conversations. In *Proceedings of NAACL-HLT*.

Gerard Salton, Edward A Fox, and Harry Wu. 1983. Extended boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036.

Nicolas Schrading, Cecilia Ovesdotter Alm, Raymond Ptucha, and Christopher Homan. 2015. An analysis of domestic abuse discourse on reddit. In *Proceedings of EMNLP*.

Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. 2018. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

Maria Smith. 1990. Aspects of the p-norm model of information retrieval: Syntactic query generation, efficiency, and theoretical properties. Technical report, Cornell University.

Richard E Smith. 2001. *Authentication: from passwords to public keys*. Addison-Wesley Longman Publishing Co., Inc.

Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. 2014. Deep neural networks for small footprint text-dependent speaker verification. In *Proceedings of ICASSP*. IEEE.

Keith Vertanen and Per Ola Kristensson. 2012. Spelling as a complementary strategy for speech recognition. In *Proceedings of INTERSPEECH*.

Robert A Wagner and Michael J Fischer. 1974. The string-to-string correction problem. *Journal of the ACM (JACM)*, 21(1):168–173.

TH Wen, D Vandyke, N Mrkšíc, M Gašíc, LM Rojas-Barahona, PH Su, S Ultes, and S Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of EACL*.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual nlu. In *Proceedings of EMNLP*.

Dit-Yan Yeung, Hong Chang, Yimin Xiong, Susan George, Ramanujan Kashi, Takashi Matsumoto, and Gerhard Rigoll. 2004. Svc2004: First international signature verification competition. In *Proceedings of ICBA*.

Steve J Young. 2002. Talking to machines (statistically speaking). In *Proceedings of INTERSPEECH*. Citeseer.

Lotfi A Zadeh. 1996. Fuzzy sets. In *Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers by Lotfi A Zadeh*, pages 394–432. World Scientific.

Sławomir Zadrożny and Katarzyna Nowacka. 2009. Fuzzy information retrieval model revisited. *Fuzzy Sets and Systems*, 160(15):2173–2191.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*.

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020. Crosswoz: A large-scale chinese cross-domain task-oriented dialogue dataset. *TACL*, 8:281–295.

Shlomo Zilberstein. 1996. Using anytime algorithms in intelligent systems. *AI magazine*, 17(3):73–73.

A Appendix

This appendix presents the scripted NLG prompts (see Subsection 3.2 and Subsection 4.1). For the British English locale (en-GB), see Subsection 3.2. All scripted prompts for the Polish locale (pl-PL):

- Q1: Podaj proszę swój kod pocztowy.
- Q2: Podaj go proszę jeszcze raz.
- Q3: Usłyszałam [1 2 3]. Podaj go jeszcze raz.
- Q4: Podaj teraz swoje imię i nazwisko?
- Q5: Podaj proszę swoję imię oraz nazwisko.
- Q6: Przepraszam, możesz przeliterować swoje imię i nazwisko?
- Q7: Jaka jest Twoja pełna data urodzenia?
- Q8: Podaj proszę datę urodzenia jeszcze raz.
- Q9: Usłyszałam [1 stycznia]. Podaj datę urodzenia jeszcze raz.

All scripted prompts for the French locale(fr-FR):

- Q1: Quel est votre code postale?
- Q2: Veuillez répéter votre code postale?.
- Q3: J'ai entendu [1 2 3]. Veuillez répéter votre code postale.
- Q4: Pourrais-je avoir votre nom et prénom?
- Q5: Pourrais-je avoir à nouveau votre nom et prénom
- Q6: Veuillez épeler votre nom complet?
- Q7: Quel est votre date de naissance?
- Q8: Pourrais-je avoir votre date de naissance.
- Q9: J'ai entendu [le 1er janvier]. Pourriez-vous répéter votre date de naissance.

B Appendix

This appendix presents Sankey diagrams for priming and speaker behaviour of dates (see Subsection 3.3). Transitions in the direction of priming in red; against, in blue. For the British English locale (en-GB), see Subsection 3.3 and Fig. 2.

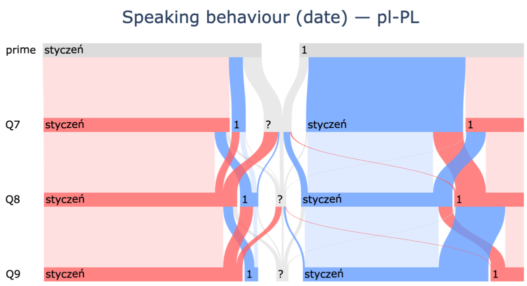

Figure 4: Polish locale (pl-PL): 85% of speakers primed with *month=name* echoed this pattern in Q_7, and only 10% of those switched later; 26% primed with *month=number* echoed and 71% later switched.

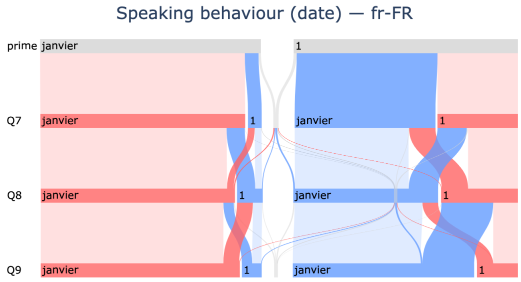

Figure 5: French locale (fr-FR): 92% of speakers primed with *month=name* echoed this pattern in Q_7, and only 9% of those switched later; 36% primed with *month=number* echoed and 67% later switched.

C Appendix

This appendix presents the target names and top-1 ASR transcriptions for all responses that employed complex spelling strategies. For the British English locale (en-GB), consult the raw data (too many examples to list exhaustively). All 10 names with complex spelling transcriptions for the Polish locale (pl-PL):

- **[Juliusz Gwara]**: **J**oanna **U**rszula **L**idia **I**wona **U**rszula **S**abina **Z**enon **G**rażyna **W**aldemar **A**nna **R**oman **A**nna
- **[Roksana Stypka]**: imię r jak **R**obert o jak **O**la kajak **K**atarzyna s jak **S**andra A jak **A**nna n jak **N**atalia a jak **A**nna nazwisko s jak **S**andra jak **T**adeusz **y** jak je t p jak **P**aulina k **K**atarzyna A jak **A**nna
- **[Nela Domino]**: dobrze imię n jak **N**atalia e jak **E**lżbieta l jak **L**uiza A jak **A**nna nazwisko The jak **D**orota o jak **O**la i jak **I**rena n jak **N**atalia o jak **O**la
- **[Róża Kochman]**: jak **r**yba **u z kreską** że jak **ż**aba A jak **A**nia
- **[Ida Heinrich]**: i jak **i**gła d jak **D**anuta a jak **A**gnieszka ha jak **H**alina e jak **E**lżbieta I jak **i**gła n jak **N**atalia r jak **R**yszard i jak **i**gła c jak **c**ebula ha Jak **C**hełm
- **[Sonia Dybiec]**: **S**abina **O**lga **N**atalia **I**rena **A**gnieszka **D**anuta **Y**eti **B**arbara **I**wona **E**lżbieta **C**elina
- **[Kalina Hus]**: **K**rystyna **A**nna **L**ucyna **I**lona **N**atalia **A**nna **H**alina **U**rszula **S**abina
- **[Elżbieta Minkina]**: **E**lżbieta **L**eokadia **Ż**aneta **B**olesław **I**lona **E**lżbieta **T**adeusz **A**nna **M**arlena **I**lona **N**atalia **K**arol **I**lona **N**atalia **A**nna
- **[Justyna Grzelczyk]**: imię J Jak **J**ustyna u jak **U**rszula s jak **S**tefan te jak **T**eresa **y** jakie t n jak **N**atalia a jak **A**nna nazwisko g jak **G**rażyna r jak **R**obert **z** jak ze mną dieta l jak **L**uiza c jak **C**ezary **z** jak zenum **y** jakie t k jak **K**atarzyna
- **[Piotr Kręcisz]**: p jak **p**ralka i jak **I**rena o jak **O**lga t jak **t**ata r jak **R**oman **k r a c z**

All 2 names with complex spelling transcriptions for the French locale (fr-FR):

- **[Timothée Samson]**: est-ce qu'on sa vie à comme Alex matrix comme **S**ophie **O**livier comme **N**athalie

- **[Constance Carlier]**: c'est con ce s'il a comme **A**lix elle comme elle est comme comme **É**milie el khomri

For the pl-PL and fr-FR locales, all listed examples are responses to Q_6 and arose spontaneously, without priming (see Subsection 3.3).

D Appendix

This appendix presents the DET plots (Subsection 4.5) for the verification task experiments (Subsection 5.2). For the British English locale (en-GB), see Subsection 5.2 and Fig. 3.

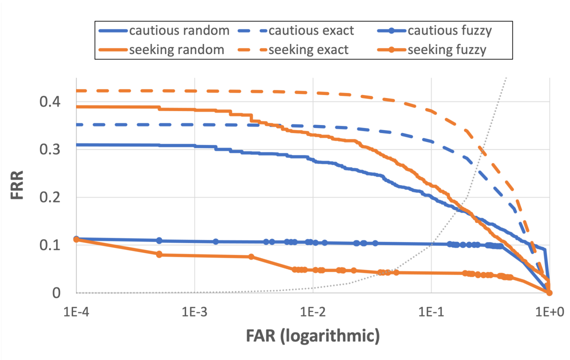

Figure 6: DET curve for the Polish locale (pl-PL)

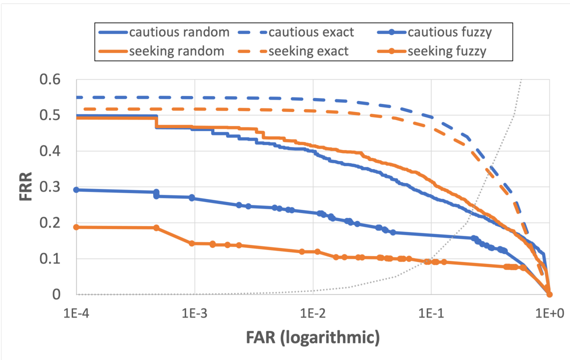

Figure 7: DET curve for the French locale (fr-FR)